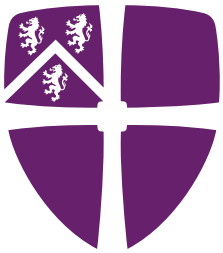
Essays in High Frequency Trading, Portfolio Selection and Oil Futures Markets

ALSHAMI, ABDULLAH

Essays in High Frequency Trading, Portfolio Selection and Oil Futures Markets



Abdullah Alshami

Business School

Durham University

A thesis submitted for the degree of

Doctor of Philosophy

December 2017

I herewith declare that I have produced this thesis without the prohibited assistance of third parties and without making use of aids other than those specified from other sources, and which have been identified as such.
This thesis has not previously been presented in an identical or similar form to any other English or foreign examination board.
The Ph.D. work was conducted from January 2014 under the supervision of Prof. Dr. Julian M. Williams at Durham University.
Abdullah Alshami Durham, United Kingdom

Acknowledgements

This thesis is the result of my research as a doctoral student at the Durham University Business School, Durham University between January 2014 to December 2017. During my research.

First of all, I would like to sincerely thank mother, the first believer in me, she grow me up and gave me all the love and the instruments that can help me to fulfill my destiny. As she always says for the past 35 years " I have faith in you son".

Also, I would thank my wife, she supported me in strong manners, helping me to rise my lovely daughters (Wadha, and Monera) the joy of my life, while I am away from them.

Moreover, for Professor Julian Williams, the great man the great supervisor. I can not find any word in the dictionary that enough to give justice for what he gave me for the past 4 years in my Phd journey.

He was the greatest ever man who supported me, I am in debt for him for the rest of my life, for what he taught and, done for me.

I am extremely glade and proud to be one of his students.
As I always say Professor Julian the legendary supervisor.

I would also thank my colleagues, Dr. Jing Nee, and Dr. Pongsutti Phuensane, who were helping me in my Phd. Journey.

Abstract

High frequency trading (HFT) requires a detailed analysis of the quote structure of the continuous limit order book in order to correctly derive viable arbitrage strategies. Traders can manipulate order books by submitting and retracting 'spoof' orders at various levels of the order book by introducing, quote volume at or above (below) the best ask(bid). However, the limit order book data for heavily traded financial instruments presents an almost unique problem to the econometrician interested in constructing high frequency measures of liquidity impact over and above the inside spread. A single month of data for an individual maturity of an activity traded futures contract, in our example light crude, can easily exceed 10 Billion bytes of data, even when stored using the single precision floating point format.

In this thesis we conduct a large scale analysis of the West Texas Intermediate (WTI) futures contract across the 120 simultaneously traded maturities for five levels of the order book from 2008 to 2016 sample at the continuous limit. Using this very-large data-set we estimate a new form of realized vector autoregression and derive the impulse response functions useful in building a HFT strategy. we show that for WTI futures a speed of execution of the order of 100s of milliseconds is needed to fully exploit a false quoting strategy designed to systematically unbalance the order flow. Furthermore, we demonstrate that viable strategies can be built by spoofing up to three levels above the inside spread.

A second part of the thesis involves creating new bootstrap routines to extract meaningful composition data to generate factor pricing models from high frequency data. The key element of this analysis is in understanding the eigendecomposition and subsequent principal component analysis to extract factors from the data. our bootstrap is new and we provide an analysis of power and consistency in correcting bias in the estimation of the eigenstructure and hence evaluating the optimal number of principal components within the data.

Nomenclature

As far as possible a single coherent notation was used. Nevertheless, sometimes a certain variable may have different meanings.

A	Ask price
B	Bid price
P	Future price
V	Volume
n	number of trader
a	time interval
X	Short possition
$\{n_p\}$	sub index
$\{\alpha_p\}$	index arrival time
$\mathscr{A}^I$	An arithmetic for computing market statistics
$\mathcal{LN}$	log normal distribution
$\mathcal{N}$	normal distribution
$\mathcal{G}$	Gamma distribution
$\mathcal{W}$	Wishart distribution
$\mathbf{v}_t$	Volume victor
$c_{i,t}$	relative quote volume concentration index
$q_{l,t}$	quote volume concentration index

$C(L)$	polynomial lag operator
Π	matrix of coefficients of interest, of order Π_i, for $i \in \{p, q, r\}$, with generic element $\tilde{\pi}$ and estimator bias $\tilde{\psi}$
$\hat{\Gamma}_h$	h-th order auto-covariance matrix
$\tilde{W}$	Newey and West demonstrate
K	Realized Kernel
$(.)_r$	Integer indexations in subscript lowercase latin, normally indexed over their adjacent roman capital, example: $r \in (1, \ldots, \mathrm{R})$ unless otherwise stated.
$\hat{A}$	Unbiased estimator of A.
$\mathbb{E}_a$	Unbiased expectation under condition c.
$\wedge$	Wedge product of two vectors, solving .
$A_\bullet$	Place holder for a function or operator, normally operating within a unit circle.
$A_\bullet^{p,q}$	Place holder for the derivetive/anti-derivative of a function or operator.
$K(.)$	Kernel function for kernel regressions.
w_Y	Sample spectrum of random univariate process Y.
$\tilde{W}_T^*$	Sample spectrum of multivariate process, for data with sample size T.
θ_i	Parameter vector as part of a collection $\theta = \{\theta_a, \theta_b, \theta_c\}$.
$\mathbf{\Psi}$	Autocross-covariance spectrum and estimator $\hat{\mathbf{\Psi}}$.
$R(.)$	Kernal estimated realized variance-covariance and cross covariance matrix..
$\int_0^T \mathscr{F}(.)ds$	Lebesgue integral with respect to state-space s.
$\mathbf{F}$	Matrix of factors.
$\mathbf{\Lambda}$	Matrix of factor loadings.
$\mathbf{A}$	Matrix of eigenvectors, usually of dimension $h < n$.
N, T	Number of observations, generically and time series.

n	$N-1$.
Σ, Ω	Generic 'true' variance covariance matrices.
$\otimes$	Kronecker product.
etr	Exponential trace function, $\exp(tr)$.
x_a	Marginal random variable indexed by a.
etr	Exponential trace function, $\exp(tr)$.
$\widetilde{IV}$	Integrated Variance-Covariance Estimator.
H_a	Null hypothesus indexed by a, where $a = 0$ is the null.
H_a	Null hypothesus indexed by a, where $a = 0$ is the null.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Direction of the thesis

The primary focus of this thesis is in developing tools that can be used to analyze high frequency pricing data and in particular high frequency data generated from crude oil futures data. Crude oil is one of the most actively traded futures, the primary benchmark of interest for this thesis is the West Texas Intermediate (WTI) contract.

In this thesis we will develop a new set of tools that we will then apply to oil futures data and with comparison to other more commonly used data sets such as the S&P 500 cross section. These tools are designed to account for the peculiar set of properties prevalent in data of this type, notably excess volatility, variable auto-correlation and cross correlation that switches sign and considerable degrees of skewness and kurtosis in the underlying random variation of price changes.

To this end we develop two main tools, with a series of sub-features. The first

is designed to take advantage of a unique data set that is the complete limit order book recording the trades and quotes within the market.

The properties of the limit order book are documented in Chapter 2 and in Chapter 2 we motivate the construction of the estimator of a vector regression model given the underlying properties of the data. As of production of this thesis, we are quite certain that this is the first study to look at the price evolution of the entire limit order book as a multivariate process. To this end we specify a new type of vector autoregression in Chapter 3 that builds on the principles of the approach outlined in Newey and West [1986] and applies this to a case with a parametric form driven by a complex the disturbance structure. In the thesis we outline the specific structure of the estimator and then run simulations to check for bias and consistency. Once this estimator is checked, the estimator is used to parameterize a model of the limit order book, where the order flow imbalance is used to forecast the mid-price and spread.

The results of this analysis provide a clear time frame for the speed of reaction in the mid-price of shocks to the flow of new orders in the limit order book. We show that deterministic arbitrage profits can be generated by strategically placing limit orders in the order book, prior to executing a market order with pre-defined direction, this is colloquailly referred to as a 'pump-and-dump', whereby, prices at or around the best bid or ask prices are placed to artificially push the mid price and hence push the opposing best bid (up) or ask (down) on the opposite side of the limit order book. The results illustrate that high speed reactions are needed with 100 milliseconds being close to the required timing to take full advantage of changes in order-flow.

Moving on from simply looking at trade in a single contract in Chapters 4 and 5 we develop a new set of tools that look at the co-evolution of the entire futures curve, updated at a high frequency. Factor models are common approaches for describing large cross sections of data, be it a cross section of equity prices or a cross section of futures contracts with varying maturity dates. The typical approach is to use a technique such as principal component analysis (PCA) to extract a smaller number of factors from the cross section and then model the stochastic evolution of these factors to generate a model that can be used for portfolio management and/or computing hedging ratios to reduce exposures. In chapter 4 we review the existing limit theory for testing for a given number of factors and then apply this limit theory to a new bootstrap model that accounts for the disturbance structure documented in Chapter 2. The underlying approach is not new Aït-Sahalia and Xiu [2018] have proposed an asymptotic identification technique using a penalty function and information criterion to identify the factor structure of high frequency data. The approach in this thesis is to identify a Neyman-Pearson style likelihood ratio approach and we briefly compare this to the cross sectional analysis in Aït-Sahalia and Xiu [2018] using equity cross section data. Once the classical theory is outlined and the critical components extracted, the bootstrap is specified and shown to be consistent with the asymptotic theory under normality. we then take advantage of the bootstraps better sample properties to identify a factor structure and then back test the efficiency of the hedging ratio (in terms of reduction of variance versus number of factors).

1.2 A historical account of crude oil prices, from the 1970s to 2015

Debate is still ongoing about the reasons for the volatility of crude oil prices between 2003 and 2014. On the one hand, industry professionals and politicians believe it was the financial instruments resulting from the financialization of the oil futures markets–for example, derivative and swaps that drove the rise in spot prices. On the other hand, having applied various regular and sophisticated tests, elite academics such as Büyükşahin and Harris [2011] believe there is no coherent evidence to support politicians theories. However, even with these conflicting views, it cannot be denied that oil prices reached historical highs between 2003 and 2014. In the current thesis, will use a pure financial market microstructure to analyse this surge in oil prices. The microstructure analysis will be based on high-frequency data to facilitate an in-depth investigation of oil futures using Western Texas Immediate (WTI) as a crude oil benchmark, because it has the richest data. However, first will highlight the findings of both the politicians and the academics, and will then go on to explain how to creat the analysis structure to identify fluctuations in prices.

1.3 Historical overview of oil prices: oil prices since the 1970s

The 1973/1974 oil crisis placed the rise in oil prices in focus as an issue to be monitored by countries, analysts, banks, politicians and economists. It was clear

that there was a need for the oil market to be re-governed. In essence, the flow of literature since the crisis dealt primarily with the shock of such crises and how to predict their impact on economies around the globe. In Baumeister and Kilian [2016] paper, defined a new regime in the global market for crude oil, allowing free fluctuations in response to the forces of supply and demand. Baumeister and Kilian [2016] summarize effectively how the mechanisms of the market which this thesis currently monitoring and we will look in the Oil futures in this thesis. Generally, the oil market has become sensitive to every potential shock; for example, a prediction of a cold winter in Europe or the US will raise oil prices, as will a conflict in the Middle East. For this reason, policy makers and economists try to cooperate and protect oil prices from shocks that they can predict by addressing those shocks before they cause market fluctuations.

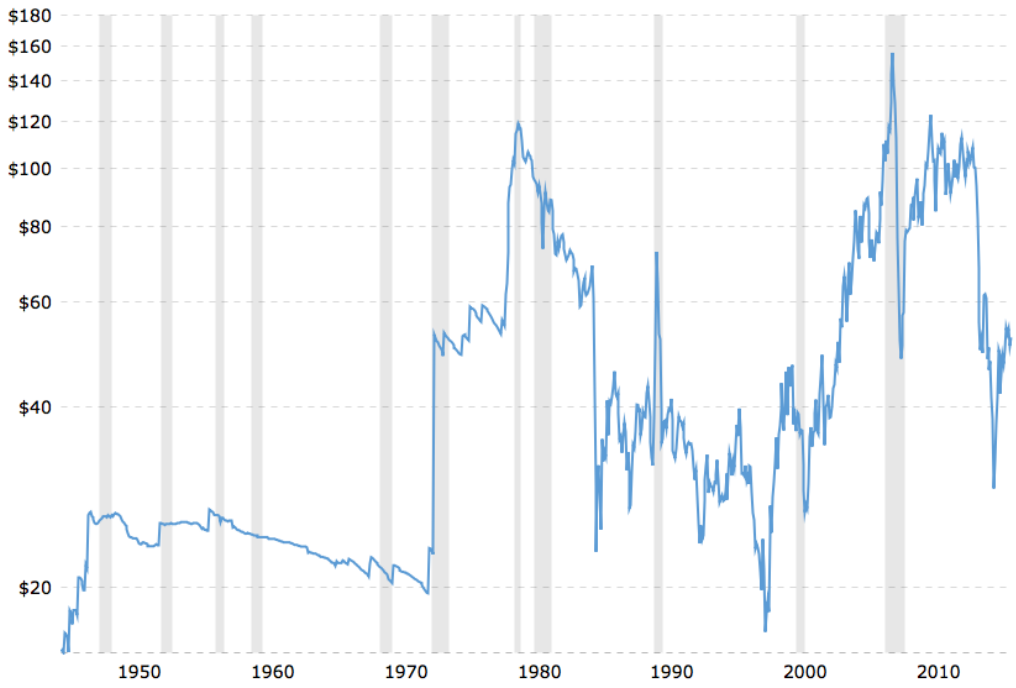


Figure 1.1: Oil prices price changing over time

Notes: Oil prices 1945–2017. Prices were stable with low volatility until the Middle East crises of 1973 and 1980. The graph also shows the price surge from 2003 to 2014, and the sharp drop from 2015. from online [2016]

1.4 The 1973 Shock

The global economy, particularly heavy industry, has been expanding for about 100 years. However, this expansion has come at a cost–namely, the need for energy and raw materials. Because these resources have largely been derived from oil, it follows that oil prices have a huge influence on the global economy. During the Middle East conflict of 1973 and early 1974, the prospect of oil supply loss from this region became a major threat to the world economy [Baumeister and Kilian, 2016] .

In this way, oil become a geopolitical card that was used as leverage against the West by policymakers in the Middle East. Basically, the oil shock of the 1970s was a supply issue that elevated the price of crude oil sharply, causing the supply curve to shift to the left along the demand curve Hamilton [2003]. However, according to Barsky and Kilian [2001], the drop in oil production in October 1973 was not the only reason for the price hike in oil. In fact, prices at the time were vulnerable to any shock for two other reasons. First, with the exception of Saudi Arabia and Kuwait, most of the oil-producing countries were reaching peak oil production. Second, the fact that countries were reaching maximum production power can be attributed to the fact that the global economy was booming, and the demand for oil was thus accelerated.

Oil prices had been fixed in 1971 for five years in the wake of the Tehran/Tripoli agreement, but the agreement failed before its expiry for the reasons mentioned above. Moreover, the prices of all global commodities increased by 75% alongside oil prices, so price increases were happening across the board. Thus, according to Kilian [2009], while the 1973/1974 shock accounted for 25% of the cause of oil price hikes, the remaining percentage was caused by other macroeconomic factors, such as supply and demand in oil sector, and global GDP healthy growth.

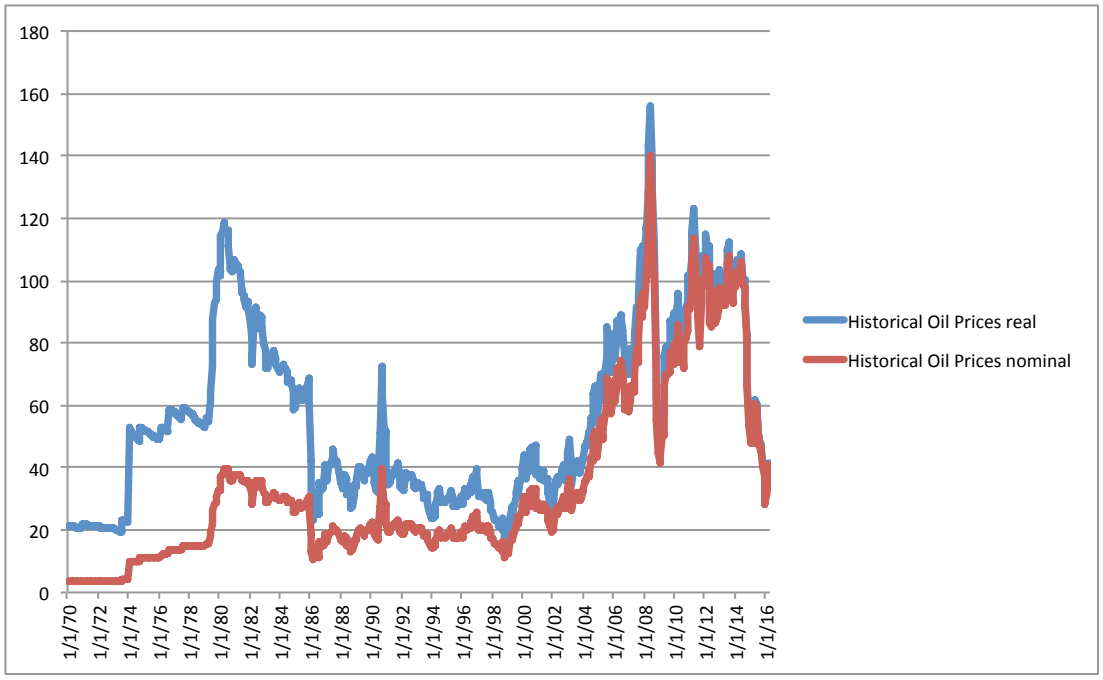


Figure 1.2: Historical Real and Nominal Oil Prices
Notes: It Shows the volatility of the oil prices in terms of the shocks in the 70's ,80's and 90's online [2016]

1.5 The crises of the 1980s and 1990s

President Jimmy Carter's government faced a high level of inflation in the US economy, and this was what led Paul Volcker, Chairman of the Federal Reserve under the Carter's administration, to raise US interest rates. Volcker's decision caused a drop in oil prices that came as something of a global shock, coming as it did from the world's largest economy. However, the shock was not destined to last very long. The first Gulf War started in September 1980, pushing oil prices to $38 from $17 Hamilton [2003]. In fact, besides Volcker's decision, there were other major factors that caused the decline in oil prices. One of those was that non-OPEC members (the UK, Norway and Mexico) joined the oil-producing countries, which caused the OPEC market share to fall in 1973 to 43% in 1980

from 53% and 28% in 1985 [Baumeister and Kilian, 2016].

Also, the global economy fell into recession in 1979. Subsequently, there was a huge drop in oil prices in 1986, which was primarily caused by a major increase in oil production by Saudi Arabia, coupled with a fall in inventory demand. In the 1990s, the Iraqi invasion of Kuwait drove prices sky high again, leading to a massive global shock [Kilian and Murphy, 2014].

In fact, Kuwait's and Iraq's oil production almost stopped amid fears that Iraq would invade Saudi Arabia, so inventory demands was also supporting prices at this time. However, when the US decided to fight Saddam Hussein (Iraqi leader), and when Saddam then retreated, oil prices reduced again in 1991 and tensions dissipated. Thereafter, oil prices remained under pressure and fluctuated until 1998, when they reched bottom at $11.

1.6 Political views of oil prices 2003 to 2014

In 2006, the US Senate investigated the surge in oil prices, because it was suspected that there was speculation in the oil futures market. As a result, a subcommittee report titled, "The role of market speculation in rising oil and gas prices" ,was published in 2006, in which it was noted that; there was substantial evidence supporting the conclusion that the large amount of recent speculation in the market had significantly increased prices. The report identified that the US regulatory system had a loophole relating to the derivatives trading market, which was being used by speculators to raise prices. Based on the above, the Senate found clear evidence of manipulation of oil prices, specifically in the case

of WTI in the NYMEX. Also, the report pointed out that the Commodity Futures Trading Commission, a financial futures regulator, had been mandated by Congress to ensure that prices in the futures market reflected the laws of supply and demand rather than manipulative practices or excessive speculation. The US Commodity Exchange Act (CEA) states that... "Excessive speculation in any commodity under contracts of sale of such commodity for future delivery . . . causing sudden or unreasonable fluctuations or unwarranted changes in the price of such commodity, is an undue and unnecessary burden on interstate commerce in such a commodity" [Baumeister and Kilian, 2016].

Thus, speculators and speculative ventures such as hedge funds, pension funds, investment banks and even oil companies were to blame for many of the fluctuations and rises in oil prices.

1.7 Academic researchs 2003 to 2014

Once reports surfaced in the media about politicians, involvement in the oil industry, academics got involved to investigate trends in oil prices. This resulted in a large body of literature, in which the most sophisticated financial tests were used to find the truth via analyses of oil futures and how they may affect spot prices. Linear Granger causality was run by Büyükşahin and Harris [2011] to obtain an explanation of traders positions – specifically, the role they play in price fluctuations. In their paper, Büyükşahin and Harris [2011] were unable to find any evidence suggesting that traders can be blamed for fluctuations, noting that there were only net positions and net position changes of speculators and com-

modity swap dealers, with little or no feedback in the reverse direction. Kilian [2008, 2009], weighed in on this position by arguing that the surge in oil prices was driven by economic factors. Fleming and Ostdiek [1999] identified some form of link between spot prices and futures in terms of oil price rises, noting that oil futures trading led to unexplainable price hikes in oil. However, this was not supported by any economic fundamentals. It was a positive volatile shock that lasted for three weeks, giving rise to a volatility surge lasting over a year. Moreover, a symmetry evidence of asymmetry was identified in the volume– volatility relationship, particularly regarding the increase in unexpected volume combined with an increase in spot market volatility, which amounted to 80% more than the decrease in volatility associated with an equivalent decrease in unexpected volume. It is quite possible that there are interesting stories behind these relationships. However, there is a negative relationship (measured by open interest) between the overall size of the oil futures market and spot market volatility, but the relationship strengthens in the case of unexpected components of open interest. Thus, the oil futures market provides depth and liquidity to the spot markets Fleming and Ostdiek [1999], which we will discuss in our analysis later in this thesis in chapters 2,3, and 4, because an understanding of futures market behaviour is crucial. To sum up, Fleming and Ostdiek [1999] did not find concrete evidence of oil futures shifting spot prices, except in the first year that deviations began in the oil market. However, although their tests showed some trends of a mitigating effect on volatility across time periods, coupled with a positive relationship between futures volumes and volatility, the lack of resources in 1999 in terms of data and technical tools did nothing to alleviate their specific concerns

regarding volatility. The above was merely a simple test of the relationship between spot prices and oil futures. Thus, Beidas-Strom and Pescatori [2014] used a more sophisticated test, namely, a sign-restricted structural vector autoregression (SVAR) test. Of course, by 2014, researchers were enjoying the benefits of easily accessible data and much better technology than in 1999. Beidas-Strom and Pescatori [2014] performed some extra analysis on the Kilian and Murphy [2014] SVAR framework (2013) which they use from Kilian and Hicks [2013], applying more restrictions to it and using economic theories. Their null hypothesis was that only oil market fundamentals (or related news) can induce low-frequency movements in oil prices, whereas temporary under–pricing or mispricing in the futures market does not contribute to low-frequency prices. They also noted that mispricing of oil futures and global interest rate shocks can be classified as speculative demand shocks, which can cause price shifts by shifting inventory demand. Thus, Kilian and Murphy [2014] had reasonable cause to believe that financial speculation may have an effect of between 3% and 22% on short-term volatility, whereas speculation shock has an effect of between 10% and 35%, and demand shock shows a greater impact of between 40% and 45%. Kaufmann [2011] expressed another speculation-related view, believing that speculation cannot heavily influence prices, even excessive speculation, because arbitration positions in buying and selling will change simultaneously. However, some noise traders can shift prices in short and long positions, but only by a small fraction. Kaufmann [2011] believes that market fundamentals have the most decisive effect on prices, and in his study, he improved on the Kilian [2009] model for global activity, which was basically a vector autoregression model. He stated that its the demand which

made the surge of 2003–2014. While Kilian [2009] noted that shifts determine the component of real global economy activity that drives commodities in global markets. However, Kaufmann [2011] found that the increase in global demand for crude oil increased from 74 million barrel per day in 2003 to 87 million barrel per day, and it was this demand, rather than speculation, that caused the surge in oil prices. Fattouh, Kilian, and Mahadeva [2012] carried out in-depth research on the period between 2003 and 2014. In their paper, they noted the lack of definitive evidence in support of the theory that speculation in oil markets drives oil prices, whether up or down. In fact, they attributed these trends to macro economic aspects, the normal inventory influence on oil prices, and the geopolitical effect. Also, they believed that futures prices do not generally have a major effect on oil prices. Essentially, our contribution in this thesis is to carry out an even more comprehensive analysis than that of Fattouh, Kilian, and Mahadeva [2012] by using high–frequency data rather than low–frequency data to ascertain whether the above interpretation will hold.

In the third quarter of 2008, when financial crisis struck the global economy, crude oil prices experienced a significant drop from $134 to $39 for the barrel by February 2009. However, they recovered within a shorter time frame than expected, supported by continued oil consumption as part of the global GDP, which was hardly affected by the financial crisis. This indicated that there was still a demand for oil, which forced the market to use that window to recover from the shock. Later, geopolitical events in the Middle East and the chaos of Middle Eastern government regimes such as Libya, had a huge impact on crude oil prices reaching their highest historical levels [Fattouh, Kilian, and Mahadeva,

2012].

The third quarter of 2014 brought with it a new chapter in crude oil prices, with aggressive price drops that affected oil producers. Some analysts believe that there is now a hidden war between the old producers and the newcomers to the market; (for example, the Shale oil companies), with the more seasoned players attempting to push prices down in a bid to make the shale oil less profitable. The breakeven point of shell oil was $86, and prices ranged between $81 and $70. The new techniques and advanced technology introduced to the shale oil production processes drove the breakeven point down to approximately $45. The older producers did not react by minimizing their production, which resulted in over–supply in the market and prices dropping below $40. As a result of this price war, the traditional oil–producing countries, particularly in OPEC, began to face difficulties in terms of their 2015/2016 budgets. In response, they introduced austerity plans and actions on the ground in order to right their budgets and avoid deficit. However, they did not introduce plans to cut their supply to the oil market, and their prices came under huge pressure as a result, with market shares going to either competitors the new rival shale oil producers. Also, in a further new development, shale oil made the U.S the biggest oil producer in the world at 13 million barrels per day.

1.8 Oil market microstructure: oil expectations

A key issue in finance is forecasting, and that also applies to the oil market. In terms of arbitrage and hedging, oil futures are essential tools. Using oil futures

as a measure of market expectations "could be valid if the risk premium, defined as the compensation arbitrageurs receive for assuming the price risk faced by hedgers in the oil futures market, is negligible" [Hamilton and Wu, 2014].

However, This assumption is questionable' Hamilton and Wu [2014] . Fattouh, Kilian, and Mahadeva [2012] put forward a strong argument regarding oil futures and spot prices, noting that in their VAR model, futures had no effect on spot prices, despite the number of participants and schemes in the futures market,for example, hedge funds, pension funds, insurance companies and investors. In the case of all of these, their involvement in the market is facilitated by many instruments, including options and index funds, and they also use technology excessively with those instruments. In concluding their papers Fattouh, Kilian, and Mahadeva [2012] and, Hamilton and Wu [2014] , all of the above authors left the door open for other researchers to ascertain whether or not financial engineering has a major role to play in oil futures.

In a Bank of Canada discussion paper, Alquist and Gervais [2013] examined the role of speculation between 2003 and 2008. According to Buyuksahin, Haigh, Harris, Overdahl, and Robe [2008], the number of noncommercial firms trading in oil futures increased from 20% to 40% in 2008, with suspicions that the non-commercial firms in NYMEX are therefore to blame for excessive speculation leading to oil price rises. Alquist and Gervais [2013] reject the null hypothesis that changes in oil prices do not predict changes in the net positions of commercial and noncommercial firms. Also, when they ran their Granger causality, they rejected the null hypothesis that changes in the positions of both noncommercial and commercial firms do not forecast subsequent changes in prices. In addition,

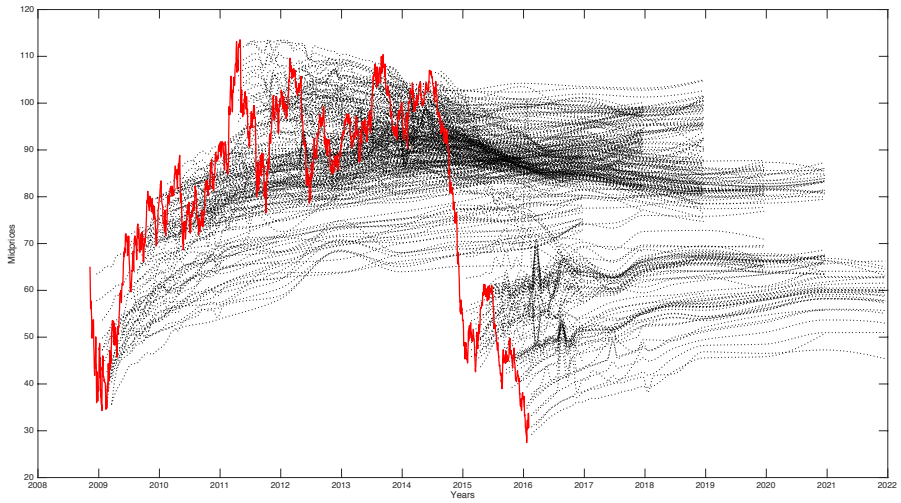


Figure 1.3: Oil futures from the thesis data

Notes: This figure illustrates the end-of-day WTI Crude Futures, form our High frequency data provided by Thomsons Ruters, in micro second. Prices from the nearest delivery [black unbroken line], which proxies the spot price, to the maximum maturity traded from July 2008 to Feb 2016.

they rejected the null hypothesis that changes in prices help to predict changes in firms' positions, based on the Granger causality test results. Thus, it was clear to them that the real interest rates, coupled with the global supply and demand conditions that had been enhanced in the period from 2003 to 2008, as well as East Asian growth, were the main drivers of oil price inflation. However, speculation may have played a modest role in these price rises.

1.9 Speculation effects

To out this thesis as following. In chapter two will apply the high-frequency trading, using brand new market microstructures analysis, with enormous data, (about 10 terabytes) data. Here, will focus on the speculations and the speculative

effects in the oil futures market for West Texas Intermediate (WTI), and the techniques of the quoting, and how big firms use the algorithms to make trading in the markets and their techniques.

Also,will use the limited order book analysis with market depth to capture the manipulation in the trading for oil futures (WTI). By collecting the tick–by–tick for oil futures for (WTI) from July 2008 to Feb 2016. For the analysis will specially created a Vector Autoregression to analyze the collected data empirically by milliseconds. The mid–price between bids and ask (for the limited order book) during the period from July 2008 to Feb 2016 oil futures for (WTI), then will examine how the shifts in the prices in the 5 levels for each bids and ask , in terms of volume, and prices all in milliseconds. We will use both univariate and multivariate econometrics analysis for the order book. Then will run a Newey–West Approach, and Monte- Carlo simulation. in Chapter 3 will run Bootstrap and Impulse Response Analysis, to check how the prices respond to the shocks.

1.10 The noise, eigenvalues latent factors and the simulation

In Chapter 4 will diagnose the effects are considered as collateral effects which appears when running the tests and the data which appear, in the models in Chapters 2, and 3. High–frequency data has a great changes for variance–covariance matrices distribution see Aït-Sahalia and Xiu [2015] , so there were a challenge to address the analysis with microstructure noise, to mange the eigenvalues in principal component analysis, to reduce the rank. Also in chapter 4 will use the

IID bootstrap with a simulations to extract the eigenvectors. Moreover, will use the Likelihood tests and designs new model for our hedging ratio. In addition, will show how the bootstrap and the other tests such as Bartlett's test affect the power of the test in our data. Finally we will illustrate power functions of our bootstrap results for the tests. our aim chapter 4 is to find the perfect approach under the high frequency data, in terms the analysis and maintaining the best possible ways to for asset pricing modules for oil futures.

Chapter 2

Spectral least squares for dynamic recovery of Impulses Responses from ultrahigh frequency data

2.1 Introduction

Automated trading in futures markets has been well established for several decades. However, modern high–performance computing has resulted in algorithms capable of operating at extremely high speeds, where decision making is measured in microseconds and the resulting actions are in the millisecond timeframes. Futures markets are somewhat of a case apart from equity markets, although both rely on similar trading mechanisms. First, transactions costs are very low. Second

the concept of liquidity provision, in the markets sense of the provision of a counterpart willing to buy or sell at a particular price, is not contingent on any kind of maximum quantity such as the total quantity of traded shares.

In a prescient observation made in an early compendium of research articles on futures markets, Anderson [1983] expressed concern at the ability of a small number of traders to manipulate the direction of futures in a systematic manner. Price manipulation of this type is somewhat different from the more commonly understood problems due to insider or private information. A manipulative strategy is agnostic to the underlying fundamental or 'fair' value of the contract. The approach involves constructing a systematic set of quoting within the market that results in near deterministic shifts in prices and bid-ask spreads to allow a trader to generate systematic excess returns. In general the objective of the market manipulator is to have no significant long or short position in the contract; however, over a day they may have significant positions in both long and short directions. The focus will be on the degree of determinism, at very high speed, quote mid-prices shift after shocks to the volume of quotes within the order book. To this end we will have to develop a new set of econometric tools based on realized VAR models to analyse the impact of shocks to the order flow on mid-prices and spreads.

First we will show empirically that a) trader shocking the order book at prices below the best-bid or best-ask price has between 200 and 800 milliseconds to execute their counter strategy before the noise from subsequent realignments overcomes the effect of the initial signal. Second, we will show that the level of adjustment to a shock from an additional quote; lower than the accountability

requirement on the mid–price is, on average, large enough to cross the half spread from the mid–price, such that the exchange's matching algorithm would automatically execute standing quotes on the other side of the order book (best ask in the case of a bid side quote and best bid in the case of an ask side quote). This in effect means that the returns to the strategy are only reliant on the adjustments within the side (bid or ask) order book within which the shock occurs. Given that the only limit on position size within a futures market is the size of the trading account margins, it seems reasonable to presume that the evidence presented indicates from Hendershott, Jones, and Menkveld [2011] that the futures market is probably best characterized by the weaker arbitrage condition of no unbounded profit with bounded risk (NUPBR) as opposed to the standard no free lunch with vanishing risk (NFLR) used in standard asset pricing models.

The main contribution is to unify the two predominant areas of research in high–frequency financial research: the market microstructure analysis of the order book and the realized regression literature. Our contribution is to specify the first realized vector autoregressive model designed explicitly to compute magnitude and significance of order-book impacts on objects of interest to high frequency traders, namely the mid-price volatility and the inside spread. The advantage of our approach is that it is model free from the view-point of the information structure of markets, as the regression model is extremely versatile and is evidence driven. While Hendershott, Jones, and Menkveld [2011] explained , but the this thesis providing a formal treatment of the issues of model specification from data beyond the one second time scale.

The availability of long histories of high quality limit order-book data is a

relatively recent innovation. Exchanges such as the NASDAQ, CME/ NYMEX and the NYSE now provide extensive coverage of 'market-depth' and much of this data is now available to the researcher at a reasonable cost. Nevertheless, the degree of innovation in financial markets is such that the volume of data, even for a single day, can result in formidable challenges to the econometrician, even when they are armed with modern high–performance–computeing (HPC) environments. An important target in the analysis of order-book data is a more nuanced understanding of the dynamics of the inside spread. This is the difference between the highest quoted bid price and the lowest quote ask or offered price, whose dynamics are often used as a proxy for the time variation of market liquidity. Whilst several studies have modelled the depth of the market from the order book that comprises the continuous auction no consensus on the information content of the various computable measures has emerged.

Popular models used to analyse such data are implemented using algorithms that have complexity ranging from low polynomial (P) to non-deterministic polynomial (NP) time is to distribute the data analysis across a high–performance cluster. However, each individual node has capacity restrictions that (a) increase the overhead of data moving across the cluster due to dependences and (b) the number of 'within-node' calculations that can be performed, in order to compute the desired component of the algorithm.

The contribution of this chapter is threefold. First, we will document in detail the mechanics of conducting both univariate and multivariate econometric analysis on order book data. Second, will consolidate a mechanism of econometric models that can be applied to order book data that provide insight into the high

frequency dynamics of volume volatility and liquidity, and apply these models to the complete history of five levels of the order book for all the maturities of the West Texas Intermediate (WTI) light crude futures delivery from July 2008 to Feb 2016 (this is 120 contracts) and the entirety of the data recorded on the SIRCA database[1] for this instrument. Third, will provide specific detailed analysis of the impact of events within the WTI light crude futures order book, specifically the impact of additional volume and traders beyond the inside quotes, on the volatility of the mid-price and the size of the inside spread. Our objective is to provide an overview of the very short term (under ten seconds) liquidity dynamics of oil futures over a long period of time and in particular to address the microstructure foundations of the widely observed maturity effect on crude oil futures.

Despite a decline in the macroeconomic impacts of oil price shocks since the OPEC I and OPEC II price shocks of the 1970s, the volatility of crude remains the most watched commodity by traders, politicians and the general public at large. A primary concern amongst the stakeholders is whether the volatility of crude oil futures prices is truly reflective of the variation in the underlying fundamentals, –see for instance Hamilton [2008] for a survey on the various theories and evidence regarding the determinants of the price of oil. At the intra-day time scale the price of crude oil is usually measured by the variability of crude oil futures prices. The two commonly traded futures contracts for the future delivery of West Texas Intermediate (WTI) light crude and the Brent-Forties-Oseberg and Ekofisk (BFOE) crude blends. WTI light crude futures are trade on the

[1]SIRCA is a provider of online services to support finance and other data-intensive research at universities, Government and financial market participants world-wide.

New York Mercantile Exchange (NYMEX) a division of the Chicago Mercantile Exchange (CME). Here we use a unique dataset of every trade and top of the order-book quotes for both the electronic trading system GLOBEX and the open outcry NYMEX pit-trades from 1996 to 2014.

Whilst this chapter focuses on the very short term interactions of the order-book for WTI light crude futures, as with any microstructure study a full appreciation of the details of the institutional arrangements and industrial organization of the futures market is critically important. The WTI settlement price is pegged to the delivery of light crude to Cushing, Oklahoma three working days prior to the 25th day of the settlement month.The WTI light crude has 120 standardised contract types for delivery each month for up to nine years. The contracts are generically coded 'CL' by NYMEX standing for light crude and then by 12 letters[1] representing each month of the year and finally by an integers 0 to 9 to represent the year of delivery. Therefore CLH0 is the March delivery for 2000, 2010 or 2020, at each maturity date the contract is rolled onto the next year ending in zero. Each individual futures contract is for the delivery of 1,000 barrels of light crude and contracts are traded via two centrally cleared mechanisms, open-outcry on the NYMEX trading-floor and the CME Globex electronic trading system and one over-the-counter, off-exchange, system CME ClearPort. On maturity the futures contracts are settled to physical delivery. Electronic trading runs from Sundays to Fridays continuously barring a 45 minute break at 5.15pm. During this time

[1]The month codes are $F, G, H^{*}, J, K, M^{**}, N, Q, U^{\dagger}, V, X, Z^{\ddagger}$ F for January, G for February, H for March, J for April, K for May, M for June, N for July, Q for August, U for September, V for October, X for November, and Z for December, for the January to December deliveries. Where $*, **, \dagger, \ddagger$ are, respectively, the quarterly deliveries for Q1 to Q4. Generally, the most actively traded contracts are the quarterly contracts, with the December "Z" contract being the most actively traded over the time period of our interest.

CME records trades and quotes through the three trading mechanisms discussed above. For GLOBEX CME the continuous auction is an electronic limit order book where bids and offers (the data is recorded as 'ask' prices) are submitted along with the required volumes. Once a trade is agreed and settled, the price and volume are then recorded. The CME order-book has some peculiarities, the most prominent being the ability to submit 'iceberg' orders. These orders have the correct price however, the volumes are disguised (usually only the 'tip' of the order is present) in order to reduce the price impact of their entry into the order stack.

Within the literature of this thesis, the most similar study to this one is Bessembinder, Panayides, and Venkataraman [2009] who provide evidence from the Base de Donnees de Marche database for 100 Euronext-Paris firms in January 2003. They find that the ability to disguise the size of trades profers asymmetric benefits across the various types of traders. For instance,patient traders tend to benefit from this hidden liquidity at the expense of those trading faster more aggressive positions. An explanation for sudden discontinuities in the price process (usually referred to ask jumps) is based on sudden aggressive repositioning of the order book. This repositioning, is possibly due to more information entering the market or because of some market dis-functionality that is suddenly corrected.

The aim is to estimate a fully specified dynamic model of the order-book and then use this model to simulate shocks to the depth of the market as measured by the volume and number of traders at each level.

The remainder of this chapter is organized as follows. In §(2.2) we review related work on oil futures pricing, order book characteristics, high frequency

trading and realized covariance estimation for regression analysis. Using this as a foundation we build an empirical model of the order-book in §(2.3) and provide some evidence using a small Monte-Carlo study on the consistency and efficiency of our chosen realized estimators. In §(3.1) we will apply the empirical application and, will review the dataset and provide a detailed review of the statistical properties of the light crude oil future order book. while in §(3.2) will present a summary of the key results, including the relative impact of high speed quoting on the inside spread and mid-price volatility, the key price benchmarks.

2.2 Literature review

At longer time-scales than intra-day the trading mechanisms driving futures prices has occupied the finance literature for a considerable time. A very early contribution is in 'The Industrial Organisation Of Futures Markets' Anderson [1983] provides, a compendium of work covering the trading structure of derivatives markets and in particular an early version of the Kyle [1985a] paper that applies a market manipulation model to a general futures market. From the viewpoint of oil futures, more specifically, Bohi and Toman [1987]and, Overdahl [1987] provide an early treatments of empirical analyses on futures trading and general market conditions. This has been further analysed in work such as Huang, Masulis, and Stoll [1996]; Moosa and Al-Loughani [1995]; Peroni and McNown [1998]; Quan [1992] who find various degrees of predictability at the daily, weekly and monthly frequencies. The high degree of heterogeniety across results found in the early literature is indicative of (a) the appropriateness of the methodology and (b)

the speedy evolution of change in the approach to trading oil futures from their inception in the early 1980s.

More recent work on the oil futures market, (see for instance Bhar and Lee [2011]; Switzer and El-Khoury [2007]; Wang, Wu, and Yang [2008]) have found that excess quadratic variation in futures markets and heightened levels of time-series persistence in prices and volatility are commonly related to the liquidity structure of the futures market. Liquidity in this context refers to market depth, and measurement of this concept has again been the subject of continuous discussion, –see for instance Fattouh and Mahadeva [2014]; Hedi Arouri and Khuong Nguyen [2010]; Nakajima and Ohashi [2012] for a comprehensive analysis of participation and liquidity of oil futures markets and the relative time variation.

Whilst the results from the empirical literature on oil futures prices is quite disparate, a common theme that harks back to the foundational comments in several of the contributions to Anderson [1983] and in particular the contribution of Kyle [1985b], is that the liquidity structure of futures markets under certain conditions lends itself to manipulation. If acknowledge that this is a possibility then the results on the lack of predictability between oil futures and oil spot physical delivery prices, (see Lee and Zeng [2011]; Silvério and Szklo [2012] for some recent results), are more easily rationalized. Put simply, oil futures appear not to fulfil their primary purpose in hedging spot price risk and this may go someway to answering the question posed in Hamilton [2008], as to whether the structure of the market is prone to systematic deviation from the efficient price due to the technology of the continuous auction.

Detecting the causation mechanism for inefficiencies at a time scale slower than daily is almost impossible. Overlapping shocks and the overlapping generations of traders across the maturity of futures contracts result in an entanglement of effects. Therefore, we can only really able to design econometric methodologies that detect the resulting ineffectiveness of the futures contracts in providing a 'rational' prediction of future spot prices. This detection of the effects of trading behaviour has dominated the prior literature in this area. Our approach is to look at the very high frequency domain of the limit order-book itself.

When Anderson [1983] was published, the ability to analyze data at the transaction level was limited by (a) technological constraints and (b) by data availability and storage. However, recent research has begun to utilize new innovations in data analysis to provide insight at the transaction level. For instance Bessembinder, Panayides, and Venkataraman [2009] provides a regression analysis of hidden liquidity in electronic markets, finding that trading strategies can impact aggregate market liquidity. Looking at technical trading at very high frequencies, Gsell [2009]; Gsell and Gomber [2009]; Hendershott, Jones, and Menkveld [2011]; Hendershott and Riordan [2009] provide a variety of examples to illustrate the importance of speed in trading and the advantage that market participants with substantial technical advantages have in driving the market in their preferred direction. The extent to which market participants can shape the market by ultra-high frequency quoting has resulted in a series of studies that look specifically at designing trading mechanisms to limit the effectiveness of these strategies, see for instance Budish, Cramton, and Shim [2013],and Easley, Hendershott, and Ramadorai [2014] for insightful comment in this area.

The majority of the current literature utilizes the traditional mechanism of techniques to interrogate the impact of speed and trading behaviour on the liquidity of markets. For instance, Hendershott, Jones, and Menkveld [2011] utilize vector autoregressive (VAR) models to measure the impact of algorithmic trading on liquidity and Bessembinder, Panayides, and Venkataraman [2009] utilizes standard single equation regression models to analyze the impact of order exposures on liquidity spreads.

High–frequency data at intraday timescales presents some formidable challenges for standard regression techniques. In response to this, a second strand of literature has emerged that specifically addresses some of these issues. Measurement of quadratic variation and covariation is of primary importance to the estimation of most types of regression analysis. Early work on measuring the realized variation in asset returns from high frequency data has a very long history. However the modern treatments mostly draw their roots from Andersen, Bollerslev, Diebold, and Labys [2001], who estimate long time-series of daily volatilities from high–frequency data. From this foundational work three important sets of results for regression analysis have been provided in [Barndorff-Nielsen, 2002; Barndorff-Nielsen and Shephard, 2004b; Hayashi, Yoshida, et al., 2005].

One of the substantive contributions is the move from calendar time to update time and the importance of this transition in reducing the complexity of dealing with data contaminants that are artefacts of the trading process. A key approach has been to use spectral methods to aggregate autocovariances to provide estimates of contemporaneous quadratic variation and covariation see [Barndorff-Nielsen, 2002].

Of specific interest is in the use of multi-timescales and various aggregation kernels to reduce contamination of estimators and consistently identify systematic covariation.

A systematic treatment of the extensibility of realized estimators from the univariate to the multivariate case found in Barndorff-Nielsen, Hansen, Lunde, and Shephard [2008, 2009a, 2011] who provide a detailed analysis of several kernels useful in ultra–high–frequency analysis of both trade and quote data. In an important contribution, Jacod, Li, Mykland, Podolskij, and Vetter [2009a] provide an extensive treatment of the theoretical properties of microstructure noise. Further work in Zhang [2011] provides a comprehensive treatment of correlations at high frequencies. Of key interest in trading futures contracts is the general breakdown in correlations as the speed of update hits the continuous limit. In this case, we can begin to see that the ability to trade quickly (at the millisecond and microsecond timescale) may provide substantial profits as the evolution of spot and futures prices decouple. It is useful to now review some core concepts from the legal literature on what constitutes the legal exploitation of a technological advantage and the illegal use of technology to artificially create arbitrage opportunities.

Budish, Cramton, and Shim [2013] give an example of a trading strategy specifically designed to make use of HFT to change the characteristics of the order book in a very specific way in order to manipulate the inside spread and provide a systematic trading advantage. It is worth restating the example as it provides an example of the mechanics. Quoting from the actual SEC documentation[1] the

[1]SEC. 2012. "Order Instituting Administrative and Cease-and-Desist Proceedings Pursuant to Sections 15(B) And 21C of the Securities Exchange Act of 1934 and Section 9(B) of the

detected HFT strategy was described as follows:

> "... at 11:08:55.152 a.m., the trader placed an order to sell 1,000 GWW shares at $101.34 per share. Prior to the trader placing the order, the inside bid was $101.27 and the inside ask was $101.37. The trader's sell order moved the inside ask to $101.34. From 11:08:55.164 a.m. to 11:08:55.323 a.m., the trader placed eleven orders offering to buy a total of 2,600 GWW shares at successively increasing prices from $101.29 to $101.33. During this time, the inside bid rose from $101.27 to $101.33, and the trader sold all 1,000 shares she offered to sell for $101.34 per share, completing the execution at 11:08:55.333. At 11:08:55.932, less than a second after the trader placed the initial buy order, the trader cancelled all open buy orders. At 11:08:55.991, once the trader had cancelled all of her open buy orders, the inside bid reverted to $101.27 and the inside ask reverted to $101.37."
>
> Quoted from Kirilenko and Lo [2013, Page 14:3]

Budish, Cramton, and Shim [2013] rightly concentrate on the speed of execution of the algorithm, but an equally interesting question is how the algorithm was calibrated to the market in the first place? The trading strategy was designed to increase the value of the sale of 1,000 shares from the best bid at $101.27 to the desired sale price at $101.34 by generating simulated buying pressure by rapidly submitting dummy buy orders. Once the inside bid had risen, another trader executed a bid to the original share order and completed the transaction at $101.34, at which point all of the dummy buy orders that had driven the best bid higher. The SEC viewed this execution strategy as manipulative and, as such, fined the trader and banned them from the market.

There are therefore two circumstances whereby speed advantages can create trading profits. The first is by, taking advantage of exogenous circumstances such as deviations in parity between two traded assets with known exchangeability, see for instance Budish, Cramton, and Shim [2013] who look at the S&P500 futures

Investment Company Act of 1940, Making Findings, and Imposing Remedial Sanctions and Cease-and-Desist Orders." Administrative Proceeding, File No. 3-15046. September 25, cited in Kirilenko and Lo [2013, Page 14:3]]

and depository receipt parities. In this instance variation is assigned to exogenous updates from one asset to another. Therefore the HFT aspect is reactionary to the price update and the advantage of speed is in taking advantage of breakdowns in correlations between assets with parity price conditions attached.

The second circumstance is the situation when the trading strategy is designed to proactively generate conditions for which the ability to execute trades and quotes at high speed can generate excess returns with vanishing risk.

Causality is therefore the most appropriate method of testing for the capacity to generate such conditions. The legal literature, (see De Charms [2013]; Guttentag [2012]; Schaffer [2010]; Wright [1985]) refers to the concept of a 'necessary element of a set of conditions jointly sufficient for the result' termed NESS. We can consider Granger causation from parts of the order beyond the inside spread as providing a trading strategy that permits the 'forcing' of the arbitrage conditions rather than playing a reactive role. In a further legal context Hart and Honoré [1985] defines legal causation by a person as an action that is part of the causation set driven by a deliberate or intended act. It is to this definition that our empirical approach is primarily directed.

Consider the following causation mechanism from the preceding example: in a thin market a sudden addition of volume and, traders to one side of the market bid(ask) can force an adjustment to the inside spread and, the degree of variation in the mid-price. If statistical causation provides sufficiently prescience then a Budish, Cramton, and Shim [2013] arbitrage strategy can be applied in reaction to the adjustment in the volatility and inside spread. we will now outline the approach to calibrating the impact of order book events ('caused' by single quote

impacts on the bid and ask) on the inside spread.

2.3 The empirical approach

Our general specification presumes that the log valuation process is generated by a continuous auction. The objective will be to compute 4 metrics of interest in a joint vector; these are the mid–price return r_t, the update speed, τ_t, the bid–ask spread by order book level $s_{i,t}$ and the oil future volume balance $v_{i,t}$ by level, with each level indexed by i up to 10 levels i=1,2,....10. It is convenient to normalize each metric so that they have a zero mean and unit variance. These four objects effectively summarize the stochastic structure of the order book and provide information on price direction, liquidity, speed and buying/selling pressure. Our main interest will be in determining the precise impact of the order flow shocks (something that as a HFT can be potentially initiate) on the mid-price and hence whether the strategy outlined in the example in §(2.2) could be replicated for oil futures, a more liquid market. The inclusion of the update speed τ_t as an endogenous variable is a unique feature of this modelling approach in the sense that speed is now directly anticipated rather than presumed to be a highly–sophisticated stochastic process.

Therefore the process presume that the change in price is proportional to the buy and sell pressure at any given time. Recall, that our interest here is only in the limit order-book and not in actual trades, which are assumed to be an artefact of the order-book. Let the tuples $\{P_n^A(a), V_n^A(a)\}$ and $\{P_n^B(a), V_n^B(a)\}$ where P denote the price, (V) is the volume, (A) is the ask, and (B) is the bid, be the

continuous time realizations of the $n \in \mathfrak{N}$ traders (it is sometimes easier to think of n as account connections rather than traders) valuations, where $\mathfrak{N}$ is the total pool of traders in the market. The a is in the interval $[0, 1]$ and represents a fraction of a day trading. As this is a futures market there is no actual physical upper limit on the positions. Additionally, traders are assumed to be able to submit both buy and sell orders simultaneously. For notational compactness we will assume that each trader can make one buy and one sell order at any given time. However, there is no real loss of generality for our purposes as we presume that individual trader strategies are exogenous. Therefore, for the n traders their oil futures aggregate short position in the order is $-X_n(a) = P_n^A(a)V_n^A(a) - P_n^B(a)V_n^B(a)$.

2.3.1 Order types

For a NYMEX futures contract traded electronically at any given instant a trader can choose to submit the following types of order to the market: a pure limit order, a market order with protection or a market-limit order. While the limit order is a buy (sell) order that denotes the maximum purchase price (minimum sale price) for a contract; any portion of the order that can be matched immediately is executed and the remainder of the order remains on the limit order book until it is either executed, cancelled or expires. A market order with protection is market order with an upper (lower) reserve limit price; if a trader sets a market order with protection they still need to set an upper (lower) bound, as such the order enters the market executes at the best available prices and then once the available volume below (above) the reserve price is exhausted the remaining order sits on the order book as a limit order. A market-limit order is an order is a market

order with protection where the reserve price is set as the best ask (bid) and if the size of the order exceeds the available volume at the best ask (bid) then the remaining order sits on the order book.

2.3.2 The standing limit order book

The observed limit order book is of course ordered from the trader, as all of the executable order flow that can be matched and traded out. Therefore, the physical process is an ordered set of prices, volumes and numbers of traders on each side (buy or sell) of the order book. Let the buy side of the order-book $\mathcal{B}(a)$ be the price–ordered set of tuples of positive bid volumes such that $\mathcal{B}(a) = \{P^B_{n,j}(a), V^B_{n,j}(a)\}$ where (P) denotes price, (V) is the volume, and (B) is the bid, $j \in J^B$ is an index whereby for any two members $k, l \in J^B$, the prices are ordered such that if $k > l$ then $P^B_{n,l}(a) > P^B_{n,k}(a)$. Similarly, the sell side of the order book $\mathcal{A}(a)$ is an ordered set of ask prices $\mathcal{A}(a) = \{P^A_{n,j}(a), V^A_{n,j}(a)\}, j \in J^A$, let (A) is the ask, (P) denotes price, (V) is the volume, such that for any two members $k, l \in J^A$, the prices are ordered such that if $k > l$ then $P^A_{n,k}(a) > P^A_{n,l}(a)$. It is also worth noting that for any subset of traders with an identical price $\bar{P}$ $\{n_p\} \subset \mathfrak{N}$, where $p \in \mathcal{P}$ is a sub–index, if $P^{i\in\{A,B\}}_{n\in n_p,j}(a) = \bar{P}$, then $j = \bar{j}$, where $\bar{j}$ is constant. As such if traders have the same bid or ask price in the order book their relative position in $\mathcal{A}(a)$ or $\mathcal{B}(a)$ is identical.[1] Therefore, we can be

[1]For the purposes of theoretical exposition for our empirical methodology the order or priority of identical prices is not important; however, it is worth noting for completeness. In terms of execution of an order on standing quotes the matching system for standing quotes with identical prices the order of execution will be in priority of which standing quotes were submitted first. let $\{n_p\}$ be a subset of traders with standing quotes $P^i_{n,j'}(a)$, for $i \in \{A, B\}$ and $n \in n_p$ where j' is the smallest member of J, as such $P^i_{n,j'}(a)$ for $i \in \{A, B\}$ and $n \in n_p$ is the current highest bid (lowest ask).

consider for any subset $\{n_p\}$ to be ordered in $\{\alpha_p\}$ Then for an additional time index of arrival times $\{\alpha_p\}$ for the subset $\{n_p\}$, the exchange matches and clears the available volume $V^i_{n,j'}(a)$ in order of $\{\alpha_p\}$.

Subsequent to the definitions of $\mathcal{B}(a)$ and $\mathcal{A}(a)$, let the aggregate order flow at time a be given by

$$X(a) = \int_{\mathfrak{N}} X_n(a)\mathrm{d}n$$

Where $X(a)$ is zero when the market is in balance, such that the oil futures weighted volume of all contracts demanded by traders versus the oil futures weighted volume of contracts for sale is the same. However, this does not mean that prices are static within the order book. Let $\mathscr{A}^I : \mathbb{R}^{\mathfrak{N}} \to \mathbb{R}$ denote a volume–weighted price aggregation operator with calculation (I). Where $\mathscr{A}^I$ as an arithmetic that converts the order books $\mathcal{B}(a)$ and $\mathcal{A}(a)$ into a single representative price process. Our interest here is to understand the impact of a change in the net order flow $X(a)$ on the volume weighted average price of interest. In particular, we are interested in the composite problem, a change an individual order flow $X_n(a)$ impacting $X(a)$ and the resulting impact on the price computed from the arithmetic of interest.

The obvious prices to track is the mid–price between the best bid and best ask. We denote this as $P(a)$, and the mid–price between the volume weighted aggregate of the bid and ask sides of the order–book as $P^{\Sigma}(a)$. As the former is relatively trivial it is more interesting to start with the latter. Let $v^i_n(a) = V^i_n(a)/V^i(a)$ where $V^i(a) = \int_{\mathfrak{N}} V^i_n(a)dn$, for $i \in \{A, B\}$ then the volume–weighted

bid and ask price operations as follows:

$$P^{\Sigma,A}(a) = \int_{\mathfrak{N}} P_n^A(a) v_n^A(a) \mathrm{d}n, \quad P^{\Sigma,B}(a) = \int_{\mathfrak{N}} P_n^B(a) v_n^B(a) \mathrm{d}n, \tag{2.1}$$

and therefore,

$$P^{\Sigma}(a) = \tfrac{1}{2} P^{\Sigma,A}(a) + \tfrac{1}{2} P^{\Sigma,B}(a) = \tfrac{1}{2} \int_{\mathfrak{N}} P_n^A(a) v_n^A(a) \mathrm{d}n + \tfrac{1}{2} \int_{\mathfrak{N}} P_n^B(a) v_n^B(a) \mathrm{d}n, \tag{2.2}$$

2.3.3 Calendar versus update time

Let $t \in \{1, \ldots, T\}$ be a discrete intra-day time index, in "update time" as opposed to calendar tick time. Let $\tilde{\tau}_t$ be the calendar-time stamped quote updates, such that $\tilde{\tau}_{t+1} > \tilde{\tau}_t, \forall t \in \{1, \ldots, T\}$. This update time is for all updates on both the bid and ask side of the order-book, hence our continuous time index in the day fraction is $a_t = (\tilde{\tau}_t - \tilde{\tau}_1)/(\tilde{\tau}_T - \tilde{\tau}_1)$. Posit a random variable $\Delta\tau_t = \log(\tilde{\tau}_{t+1} - \tilde{\tau}_t) - \log \bar{\tau}$, where we set the unconditional expectation $\mathbb{E}[\Delta\tau_t] = 0$. Therefore, we set $\bar{\tau} = \mathbb{E}_{\mathcal{L}}[\tilde{\tau}_{t+1} - \tilde{\tau}_t]$, where the operator $\mathbb{E}_{\mathcal{L}}$ denotes that the expectation is 'borrowed' from the measure that ensures $\mathbb{E}[\Delta\tau_t] = 0$. For the main specification, let $\tilde{\tau}_{t+1} - \tilde{\tau}_t \sim \mathcal{LN}(\tilde{\mu}_t, \tilde{\varsigma}_t)$, where $\mathcal{LN}$ is a log normal distribution with arithmetic first and second moments $\tilde{m}_t = \exp(\tilde{\mu}_t + 0.5\tilde{\sigma}_t^2)$ and $\tilde{\varsigma}_t^2 = \exp(\tilde{\sigma}_t - 1)\tilde{m}_t$, where $(\tilde{m}_t)$ is the first moment, and $(\tilde{\varsigma}_t^2)$ is the second moment. However many other distributions are obviously valid under this set-up. The main difficulty arises when trades have essentially identical time stamps, the date is recorded to the microsecond, however, the effective ranges for the time-stamps are to the nearest millisecond. To ensure that all the time stamps are unique to ensure that $\Delta t \to$

∞, for that, we undertake the following exercise.

1. Construct the vector $\tilde{\tau} = [\tilde{\tau}_1, \ldots, \tilde{\tau}_T]$ and construct the equivalent length index vector $\mathbf{t} = [1, \ldots, T]$.

2. Construct the list of unique timestamps $\tilde{\tau}_{\mathcal{U}} = [\tilde{\tau}_j]$ and delete the time index of duplicate times to form a vector $\mathbf{t}_{\mathcal{U}}$. we use the first listed price for the unique time stamp as CME[1] informs us that the ordering is preserved from microseconds, so the potential for rearranging prices should be negligible.

3. As the time stamps are accurate to one millisecond (and indeed has find a large number of updated separated by one millisecond) therefore for a tuple of quote updates recorded at the same time index we use a simple spline to construct the interpolated time-stamps, i.e. let $\tilde{\tau}_j$ and $\tilde{\tau}_{j+1}$ be two unique time stamps and let $\tilde{\tau}_{k=1}, \ldots, \tilde{\tau}_{k=K}$, be K updates such that $\tilde{\tau}_k = \tilde{\tau}_j \forall k \in \{1, \ldots, K\}$.

4. By construction $\tilde{\tau}_{j+1} - \tilde{\tau}_j \geq 1$ms; therefore, we construct a grid such that the abscissa values for the K time stamps for interpolation are equally spaced in order of arrival between $\tilde{\tau}_{j+1} - \tilde{\tau}_j = 1ms$ at time intervals of $k/(K+1)$ms.

5. This approach ensures that each stamp has a unique update-time, but imposes that any update would have a unique time-stamp if the difference in time is greater than one millisecond.

[1] Chicago Mercantile Exchange

2.3.4 The order–book volume–weighted mid-price and bid-ask spreads

For each quote update we have the following tuple of observed data from the limit order-book $(P_{l,t}^{Quote}, V_{l,t}^{Quote}, N_{l,t}^{Quote})$, $l \in \{1, \ldots, L\}$, $Quote \in \{B = Bid, A = Ask\}$, where L is the maximum number of levels considered in the order-book. For instance $P_{2,t}^{B}$ is the second best standing price bid within the order-book at time index t. We can therefore construct two mid-prices; first, the standard mid-price which is the mid-distance between the best-bid (highest $P_{1,t}^{B}$) and best-ask prices (lowest $P_{1,t}^{A}$) cause both are the nearest to the strike price , hence $P_t = 0.5(P_{1,t}^{A} + P_{1,t}^{B})$. Alternatively we can compute a volume-weighted average price from the limit order book prices:

$$P_t^{\Sigma} = \tfrac{1}{2} \sum_{l=1}^{N} \frac{P_{l,t}^{A} V_{l,t}^{A}}{\sum_{l=1}^{N} V_{l,t}^{A}} + \tfrac{1}{2} \sum_{l=1}^{N} \frac{P_{l,t}^{B} V_{l,t}^{B}}{\sum_{l=1}^{N} V_{l,t}^{B}}$$

The returns are then computed by forward differencing the log prices $\tilde{r}_t = \log P_{t+1} - \log P_t$ and $r_t^{\Sigma} = \log P_{t+1}^{\Sigma} - \log P_t^{\Sigma}$ and then de-meaned to eliminate the constant term in the dynamic model, $r_t = \tilde{r}_t - \bar{r}$, to ensure that $\mathbb{E}[r_t] = 0$.

For the bid ask spreads we compute the simple log-difference at each time step and similarly to the time stamp case we deduct a long run average to ensure that the expected value is zero. As such we set $\tilde{s}_{l,t} = \log(P_{l,t}^{A}/P_{l,t}^{B}), \forall l \in \{1, \ldots, L\}$ and then compute $s_t = \tilde{s}_{l,t} - \bar{s}_i$, such that the unconditional expectation of the spread is zero, $\mathbb{E}[s_{l,t}] = 0$. We collect the spreads, by level, into the following vector $\mathbf{s}_t = [s_1, \ldots, s_L]'$.

2.3.5 Measuring the order flow imbalance

The main object of interest is the effect of excess quoting on one side of the order book has on the mid-price return and the speed of update to the order book. Let the order-flow imbalance be denoted by $v_t = \tilde{\nu}_{t_0}\bar{\nu}$, such that $\tilde{v}_{l,t} = \log(V^B_{l,t}/V^A_{l,t}), \forall l \in \{1,\dots,L\}$ where (v_t) is the volume over time, and the unconditional expectation of the order flow imbalance is $\mathbb{E}[v_t] = 0$. Note that we place the bid volume as the numerator and the asks as the denominator. This is to assist the logical interpretation of the strategies generated by the VAR model. When there is a positive shock to the bid-prices we would anticipate, ceteris paribus, that there will be a rise in the price. As such when we convert to the order-flow imbalance a positive increase in V^B_l increases v_t deterministically and we postulate that this will have a temporary impact on the mid-price return. The most interesting forward looking impact is not from the level one volume, but from level two and this is the commonly postulated mechanism for HFT traders in the oil market.

We can repeat this calculation for both the oil futures weight volume $\tilde{v}_{l,t} = \log(P^B_{l,t}V^B_{l,t}) - \log(P^A_{l,t}V^A_{l,t})$, $\forall l \in \{1,\dots,L\}$ and the number of contracts. There are two rationales for using just the oil futures weighted volume. First, most microstructure models such as those in the Kirilenko, Kyle, Samadi, and Tuzun [2014], family use oil futures weighted volume, so if the final objective is the direct imputation of structural parameters from these models then the order imbalance should be measured in currency and not numbers of contracts. Second, from the vantage point of generating a HFT strategy, the oil futures weighted volume provides an oil futures amount rather than a contract amount for the draw down

on their account for generating the HFT strategy. We collect the time evolution of the various levels of the order book in the sorted vector $\mathbf{v}_t = [v_1, \ldots, v_L]'$.

2.3.6 The quote volume concentration Index

For each message in the order-book the number of active accounts concurrently registering a quote is recorded. We note that in general the highest number of active feeds quoting simultaneously in a contract is not normally more than 100 buyers or sellers (recall that it is likely that the active traders will be on both sides of the market). Many standard theoretical models of market microstructure derived from the Hellwig [1996], Grossman [1987], and Stiglitz [2002] postulate that the degree of disequilibrium generated by inefficiencies (relative to classical models) will be a function of number of traders. We can either model directly the imbalance in terms of the numbers of account open at any given instant (this approximates the arrival of a new trader) or the volume of quotes per active account. We call this the 'relative-quote-volume-concentration-index' or QVCI, denoted $c_{i,t}$. The trader imbalance is computed as follows: $\tilde{c}_{l,t} = \tilde{\nu}_{l,t} = \log(N^B_{l,t}/N^A_{l,t})$, $\forall l \in \{1, \ldots, L\}$ and $c_{l,t} = \tilde{c}_{l,t} - \bar{c}_l$ such that $\mathbb{E}[c_{l,t}] = 0$. For the QVCI we compute the following indices:

$$\tilde{q}_{l,t} = \log(V^B_{l,t}/N^B_{l,t}) - \log(V^A_{l,t}/N^A_{l,t}), \forall l \in \{1, \ldots, L\}$$

and in the empirical model we use the demeaned $q_{l,t} = \tilde{q}_{l,t} - \bar{q}_l$, such that $\mathbb{E}[q_{l,t}] = 0$. The baseline specification is the joint evolution of the vector $y_t \in \mathbb{R}^n$, where $y_t = [r_t, \Delta\tau_t, \mathbf{s}'_t, \mathbf{v}'_t]'$, as a vector autoregressive process in update time. Notice,

that we are explicitly endogenizing the actual update time, therefore the timing of an adjustment in the order book is explicitly modelled by the vector of impulse response functions $\mathrm{D}[\Delta\tau_{t+s}] = \partial\Delta\tau_{t+s}/\partial(y_t - \mathbb{E}[y_t])'$. Furthermore, and arguably for a high frequency trade of more interest are the responses $\mathrm{D}[r_t]$ and $\mathrm{D}[\mathbf{s}_t']$ as these provide the direction and magnitude of adjustments to shocks to the order book in terms of direction of prices and spreads.

Of critical importance here is the notion that the time frame $\sum_{t=1}^{Y}\tau_t$ is fixed (usually a single day) and that sampling of t is driven by the realizations in update time. Hence, as $T \to \infty$ the mean update time and returns tend to zero, that is $\bar{\tau} = T^{-1}\sum_{t=1}^{Y}\tau_t \to 0$ and $\bar{r} = T^{-1}\sum_{t=1}^{Y} r_t \to 0$. However the mean order flow imbalance and spreads with tend to a constant, $\bar{s}_l = T^{-1}\sum_{t=1}^{Y} s_{l,t} \to s^*$, for a given level l and $\bar{v}_l = T^{-1}\sum_{t=1}^{Y} v_{l,t} \to v^*$. This is why our metrics for $s_{l,t}$ and $v_{l,t}$ are demeaned at the initial stage.

2.3.7 Vector autoregression

Following from this logic, our baseline specification is a linear r-th order vector autoregression where $C(L)y_t = c + u_t$, where $C(L)$ is the polynomial lag operator, c is a vector of constants and u_t is a disturbance process. In matrix notation this is written as

$$Y = X\Pi + U \tag{2.3}$$

where $Y = [y_t']_{t=r+1}^{T}$ and $X = [x_{t-1}', \ldots, x_{t-r}']_{t=r+1}^{T}$ and $x_t = [y_t, 1]$. The matrix of coefficients of interest $\Pi = [\Pi_1, \ldots, \Pi_r, c]$ provides me with an exact mechanism to identify the matrix of impulse responses $\mathrm{D}[y_{t+s}]$. Let F be the $nk \times nk$ companion

matrix for $\Pi = [\Pi_1', \ldots, \Pi_k']'$. Therefore:

$$\mathbf{F} = \begin{bmatrix} \Pi' \\ \mathbf{I} \quad \mathbf{0} \end{bmatrix}$$

Setting the upper left $n \times n$ sub-matrix from a square matrix raised to the s power as $\mathbf{F}^(s)_0$ we use the Wild representation of the impulse response function $\mathrm{D}[y_{t+s}] = \mathbf{F}_0^{(s)}$. The most useful part of this exercise is that for an estimate $\hat{\Pi}$ if $\sqrt{T}\mathrm{cov}(\mathrm{vec}\Pi - \mathrm{vec}\hat{\Pi}) \rightarrow^D \mathcal{N}(0, \hat{\mathbf{C}})$, then the asymptotic distribution of the impulse responses can be recovered using the delta method. Let $\Psi_s : \mathrm{vec}\Pi \rightarrow \mathrm{vecD}[y_{t+s}]$ be the s step impulse operator. Then the error covariance matrix is computed by

$$\sqrt{T}\mathrm{cov}(\Psi_s[\mathrm{vec}\Pi] - \Psi_s[\mathrm{vec}\hat{\Pi}]) \rightarrow^D \mathcal{N}(0, \nabla\Psi_s[\mathrm{vec}\hat{\Pi}]'\hat{\mathbf{C}}\nabla\Psi_s[\mathrm{vec}\hat{\Pi}])$$

where $\nabla\Psi_s[\mathrm{vec}\hat{\Pi}] = [\partial\Psi_s[\mathrm{vec}\hat{\Pi}]/\partial\mathrm{vec}]$. We will now illustrate an estimator that matches the properties required to compute $\hat{\Pi}$ and hence provide point estimates and confidence bands of the impulse response function. The major object of interest is to compute as exactly as possible $\partial r_t / v_{i,t}$ and $\partial \tau_t / v_{i,t}$, where $i \in \{1, \ldots, L\}$ is the level of the order book. This allows the HFT as precisely as possible to understand (a) how much a shock to the volume of limit orders on one side of the market will affect the mid-price and (b) how fast following updates to the order book will be once the shock has propagated. With this information in mind the HFT can pre-empt the directional adjustments and construct a trading strategy according to its goals.

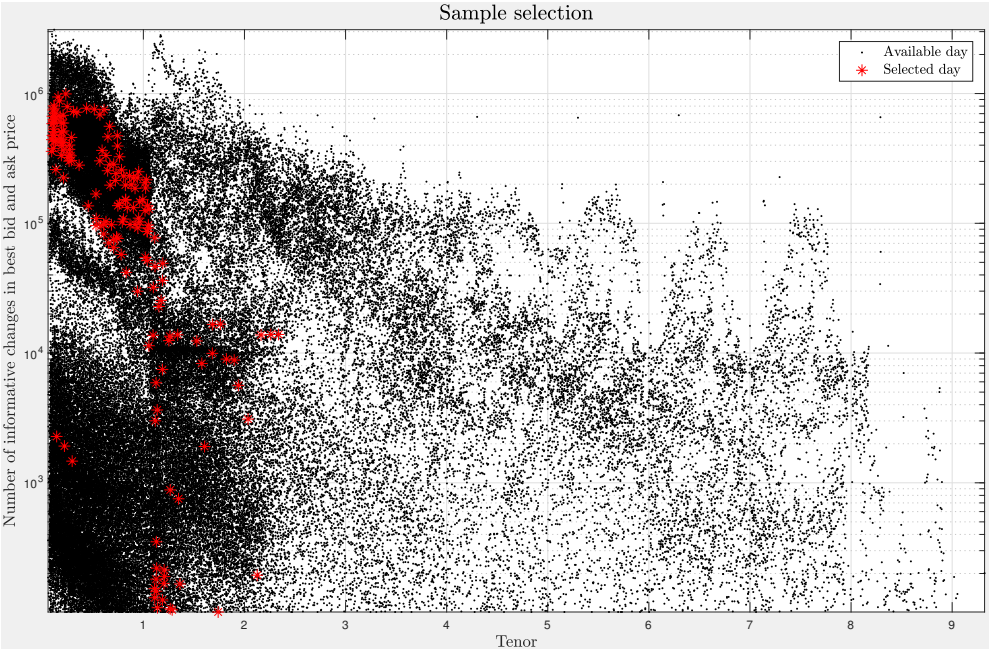


Figure 2.1: Market activity, measured by the number of informative price changes at the best bid and ask price for all contracts in the available sample and those contracts included in the VAR analysis.
Sampling method for choosing the days included within the sub-sample analysis in Chapters 3 and 4. The selected days are marked with a star.

2.3.8 Sample Selection

It is not possible to run the various models outlined in Chapters 2 and 3 over the whole sample of data, as we will utilize every informative update, hence the data management is in the form of a series of sub-samples, which are reviewed in detail in Table 2.1. In Chapters 4 and 5 we will utilize all of the sample points to create the cross sectional grid across the term structure of tenors. In Figure 2.1 we illustrate the representativeness of the sample in terms of activity at the best-bid and best-ask prices. This is the number of informative price changes at the best-bid and best-asks (level 1) in the order-book (as not every day has activity

deeper than level 1). As would be expected there are a lot more quotes than trades in the order to one to two magnitudes more changes in the mid-price than in the traded price.

We can see that the sampling takes data from around 0 to 2 years and this is when the most commonly active trading occurs. On any given day there will be activity in line with the sample and this allows us to randomly pick out trading days across the activity spectrum whilst ensuring there is sufficient information in the market depth to allow parameterization of the model. In Chapters 4 and 5 we utilize every available day in the sample, however, we do not use all of the quotes and generate time series by constructing 1 and 5 minute grids for comparison.

2.3.9 Sample Selection

In Figure 2.2 we get some interesting results from our VAR model results. We extract 6 plots to show the intraday trading among five levels (market deph) in the markets. The first 3 plots, represent 01-jul-2009. For the July 1, 2009 trading day the inside spread has 1.3 Million quote updates. Where the second 3 plots shows the November 18, 2009. There are just over 4 million updates. However the number of updates at Level 5 on July 1, 2009 is around 300 thousand as opposed to 4 million at Level 5 for November 18, 2009. The first five levels of the order book for the CLZ9 futures contract (the lowercase 'm' xrepresents the market depth indicator) for two days(July 1,2009 and November 18, 2009)over the lifecycle of the futures contract. CLZ9 is the December 2009 maturitycontract, the top plot is just under six months from maturity and the bottom plot is just under 4 weeks from maturity. The top plot presents the price at each level of

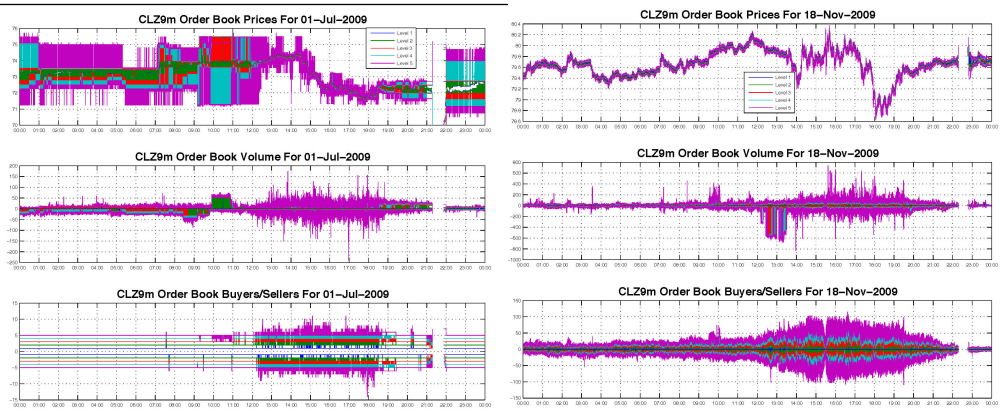


Figure 2.2: Market order book intraday and market depth for 5 levels. Top plot presents the prices for the bid and asks within the limit order book. The centre plot presents the order book volume in contracts and the bottom plot presents the number of active traders on the bid and ask side

the order book from the best bid (the lower prices) and the best ask (the upper prices); the middle plot is the volume of contracts bid or asks and the bottom plot is the number of traders at each level. Trading terminates at 22.15 GMT each day to mark each traders open interest to market.

2.4 Properties of market depth data

As this is the first study to look at market depth data for futures markets it is important to illustrate the statistical properties of the data. We will then use this preliminary analysis to motivate the use of a new estimator for a vector autoregression that fully accounts for the variational properties see Figure 2.2 as an example.

First we will take a random selection of days from the available sample. These days are documented in Table 2.1, which provides a summary of the data used in this thesis (U.S. days) . We randomly sampled 45 days of data from the complete

limit order book of all messages for those particular trading days. The selected trading days contain a number of different price levels from $38 on February 10, 2009 for the March 2009 delivery contract to $114 on April 27, 2011 for the January 2012 delivery contract.

Numbers of messages on the order-book range from eight thousand messages to over five million. The max volume refers to the size of the large transacted trade conducted that day. The largest observed trade in this dataset is 1,000 contracts at around $82 per barrel, which is a single trade with notional value 1,000 barrels times 1,000 contracts times $82 per barrel.

The 'Max Sellers' and 'Max Buyers' columns represent the number of active buyer and seller accounts connected to Globex throughout the day. We will be working in 'update-time' and the speed of update $\tilde{\tau}_j$ will be one of the endogenous variables within the equation. Update-time or business time is notoriously difficult to analyse particularly for cross-variational studies, with early documentation on the correlation problem outlined in [Epps, 1979].

The problems associated with update time effects have been analyzed in detail in Barndorff-Nielsen, Hansen, Lunde, and Shephard [2008, 2009a, 2011], who propose a series of kernel estimators to run univariate regressions and identify correlations and covariances. When reviewing ultra high frequency data such as market depth, the objective is to quantify the precise impact on the mid-price accounting for all of the dynamics within the data.

Let $\hat{\Gamma}_h = 1/T \sum_{t=h}^{T} y_t y_{t-h}$, be the h-th order auto-covariance matrix of the data set Y, where y_t is the t row recalling that $t \in \{1, \ldots, T\}$ is measured in update time. For the thesis dataset we compute the average autocovariance

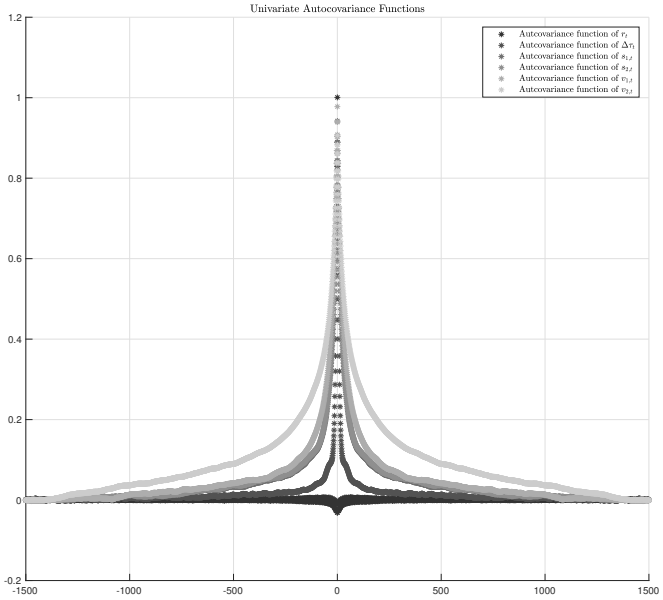
Table 2.1: Sample characteristics for the main analysis.

Code	Day	Maturity	#. of Quotes Updates	Max V.	Mid Price	Max S.	Max B.
CLF0	Dec 9, 2009	Jan-10	4 m	700	73.1-71	150	100
CLF0	Jan 3, 2009	Jan-10	L1,L2:4 m;L3:54,534;L4:6,874,L5:253	190	72-71.5	6	5
CLF1	Nov 23, 2010	Jan-11	5 m	800	82-81	300	200
CLF1	Apr 5, 2010	Jan-11	L1,L2,L3:2 m; L4:5,441;L5:69	140	87.8-88.5	4	4
CLF2	Nov 17, 2011	Jan-12	4.7 m	500	102-98.5	125	125
CLF2	Apr 27, 2011	Jan-12	L1,L2:5 m; L3: 83,007; L4:24,722;L5:10	100	112.7-114	4	4
CLF4	Dec 4, 2013	Jan-14	2 m	450	92-92	200	170
CLF4	May 23, 2013	Jan-14	1 m	130	87-87	5	5
CLF5	May 21, 2014	Jan-15	L1,L2: 6 m; L3: 8170; L4:8; L5:0	120	97-97	3	6
CLF5	Dec 6, 2013	Jan-15	L1:202;L2:171;L3:59;L4:4;L5:0	13	93.2-93.2	5	1
CLF9	Dec 10, 2008	Jan-09	3 m	500	42.3-44.1	70	79
CLF9	Nov 30, 2008	Jan-09	8,000	250	55.3-53.75	30	25
CLF9	Dec 10, 2008	Jan-09	3 m	500	42.3-44.1	70	79
CLF9	Nov 30, 2008	Jan-09	8,000	250	55.3-53.75	30	25
CLG0	Jan 6, 2010	Feb-10	3 m	1000	81.6-83.2	170	170
CLG0	Jun 22, 2009	Feb-10	L1,L2:2 m; L3:34,950; L4:40; L5:0	30	73-70	3	3
CLG1	Jan 10, 2011	Feb-11	4 m	500	89.55-89.5	160	175
CLG1	Jun 3, 2010	Feb-11	L1,L2:7 m; L3,L4,L5:0	70	80-80.5	2	3
CLG1	Jan 10, 2011	Feb-11	4 m	500	89.55-89.55	175	175
CLG1	Jun 3, 2010	Feb-11	L1,L2:7 m; L3,L4,L5:0	70	80-79.8	2	3
CLG2	Jan 5, 2012	Feb-12	3 m	580	103.2-101.5	160	170
CLG2	Jun 16, 2011	Feb-12	L1,L2,L3:7 m; L4:6 m;L5:5 m	70	97.75-97.5	5	4
CLG3	Jan 9, 2013	Feb-13	2.5 m	650	93.17-93.2	250	240
CLG3	Mar 22, 2012	Feb-13	L1,L2:2 m;L3:14,830;L4:1;L5:0	30	108.3-107.3	3	4
CLG4	Jan 2, 2014	Feb-14	1.7 m	550	98.75-95.5	225	160
CLG4	Jul 4, 2013	Feb-14	L1,L2,L3:2 m; L4:7,968;L5:050	50	95.7-95.8	3	3
CLG5	Apr 23, 2014	Feb-15	L1,L2:3.5 m; L3:1.2 m;L4:161; L5:157	50	90-90	6	7
CLG5	Dec 12, 2013	Feb-15	L1:110; L2:91; L3:9; L4,L5:0	4	92.3-92.4	2	1
CLG9	Jan 6, 2009	Feb-09	3 m	500	48.5-48.3	70	80
CLG9	Dec 15, 2008	Feb-09	2 m	200	50-47.5	30	30
CLH0	Feb 5, 2010	Mar-10	4.7 m	610	73.2-71.9	140	160
CLH0	Jul 10, 2009	Mar-10	L1,L2:3 m; L3:9,397; L4:24; L5:0	55	66-65.5	4	3
CLH1	Jan 31, 2011	Mar-11	4.9 m	500	90.5-92	160	200
CLH1	Jun 25, 2010	Mar-11	L1,L2:5 m; L3:1 m; L4,L5:7,000	60	79.5-82.25	5	3
CLH2	Jan 31, 2012	Mar-12	3 m	1200	99.1-98.5	210	170
CLH2	Aug 11, 2011	Mar-12	1 m	140	84-86	7	7
CLH3	Feb 1, 2013	Mar-13	2.7 m	600	97.6-97.6	210	210
CLH3	May 3, 2012	Mar-13	L1,L2,L3:5 m;L4,L5:0	60	105.75-103.75	3	1
CLH4	Feb 3, 2014	Mar-14	2.6 m	400	97.2-96.6	210	200
CLH4	Jun 21, 2013	Mar-14	L1,L2:1 m; L3: 126,916; L4:413; L5:0	75	91.6-90.5	4	9
CLH5	Apr 17, 2014	Mar-15	L1,L2,L3:0.5 m; L4:0.4 m L5:0.2 m	110	94.8-94.7	8	5
CLH5	Dec 11, 2013	Mar-15	L1,L2:100; L3,L4,L5:0	3	91.75-91.75	2	1
CLH9	Feb 10, 2009	Mar-09	3 m	350	40-38	75	100
CLH9	Dec 31, 2008	Mar-09	1 m	370	43-46.2	15	19
CLJ0	Mar 10, 2010	Apr-10	5 m	1,200	81.3-81.75	210	230
CLJ0	Aug 7, 2009	Apr-10	L1,L2:2 m; L3:72; L4,L5:0	27	77-76	27	13

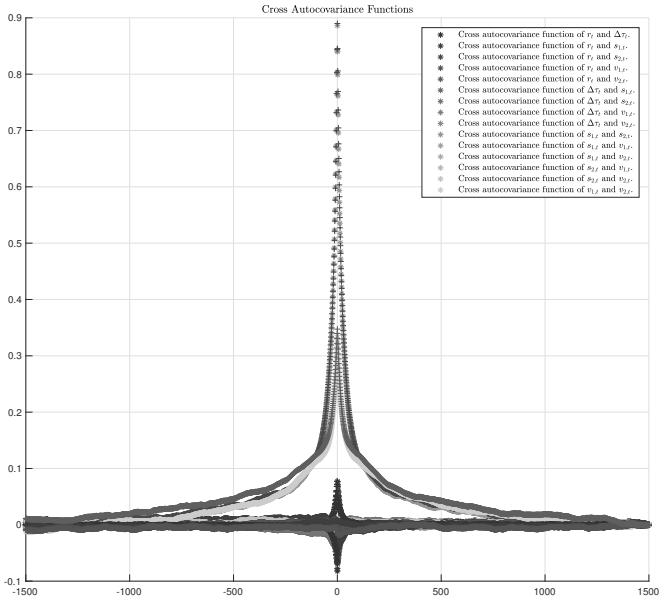
Notes: Contract code is the CME ticker code for each contract. For each code We sorted the number of trades from the day with the largest number of active quotes to the day with the fewest. We then sampled the most active day and the median activity day for comparison. Contract maturity is the date at which the contract matures to physical delivery. Number of quotes provides the number of bid-ask updates. If there is a discrepancy we use a : to present bid:ask quotes. Furthermore, if the number of quotes updates varies across the first five levels of the order book we label each activity number by level, where L = 1 to 5. Max volume is the highest total volume across the first five levels of the order book bid or ask observed on that day. Mid price provides the median price of oil per barrel for that day. Max S (Max sellers) is the highest number of sellers observed within the market for that day and Max B (Max buyers) records the highest number of buyers, both are cumulative across all levels. So a number less than 5 indicates that the highest number of buyers or sellers was less than the number of recorded levels for that day.

matrix for the days in the sample outlined above, for a range of lags from -1,500 to +1,500 (including 0) to illustrate the complex structure of the interactions. This is plotted in Figure 2.3.

Recall that the columns of Y are already normalized, hence the plots in Figure 2.3 for the autocovariance (a) and Cross-autocovariance (b) are already normalized about unity. The most obvious point to note is that the long lag and quite complex structures for all of the variables in the system, except the mid-price return r_t (darkest marker), which experiences a short term negative autocorrelation before exhibiting a slow return. However, the magnitude of the autocorrelation for the mid-price return is very small relative to the other variables, which are highly persistent. However, for the cross-autocovariances, the mid-price return and the spreads have a high absolute magnitude and very long persistence approaching 1,000 lags. Of particular notes are the cross-autocovariances with $v_{1,t}$ and in particular $v_{2,t}$. This indicates that lagged correlations between the mid-price (the response variable a HFT might seek to manipulate) and the volume balance (something the HFT can adjust) is large in magnitude if properly estimated.



(a)



(b)

Figure 2.3: Autocovariance and Cross Autocovariance functions.
Autocovariance (a) and Cross-Autocovariance (b) functions versus $\pm 1,500$ lags for the mid-price update r_t, calendar time between updates $\Delta\tau_t$, bid ask spreads for levels 1 and 2 of the orderbook $(s_{1,t}, s_{2,t})$ and the Oil Future volume bid-ask order flow imbalance for levels 1 and 2 of the orderbook $(v_{1,t}, v_{2,t})$. Recall that the cross autocovariance is not necessarily symmetrical about the zero lag.

2.5 Conclusion

In this Chapter we reviewed the literature for the algorithm trading, and examined precisely how the traders can make the maneuvers to beat the markets by the representing order book imbalance and manipulate the prices.

I have uses the level 2 of the order, see (Figure 2.2),in from the market depth data to over quoting and push the prices to their arbitrage price, all been found by the vector auto regression which been specifically developed to fulfil the contribution of this thesis as been discussed in the introduction of this chapter.

All the empirical results can been seen as an example in Table 2.1. Also the Figure 2.2 show how the traders are aggressive and can force excess returns by strategically submitting orders to the limit order bhook.

This chapter has covered the methodological part for market manipulation in terms of marketmicrostructer, and detect how and where the manipulation can be found, see Figure 2.2 and Table 2.1.

In the next chapter we will demonstrate how fast the traders needs to be in terms the high frequency trading, by applying the impulse response functions and bootstrap to recover the magnitude of expected returns using a novel estimator based on the statistical properties identified in this preliminary analysis.

Chapter 3

A Kernal VAR Estimator for Modelling Limit Order Book Dynamics

3.1 Introduction

We will now introduce a new Kernel Vector Autoregression (Kernal VAR or K-VAR) model designed specifically to analyse the limit order book at very high frequency with asynchronous data. The main theoretical construction is to build upon the Flat-Top-Kernel estimator of covariance proposed in Barndorff-Nielsen et al. [2011] and then apply this estimator to a VAR case rather than a flat covariance structure.

Autocovariance and cross-autocovariance functions are instructive for assessing the co-evolution of the variables. However they provide little or no informa-

tion on direction and causation of effects . Standard VAR analysis using OLS type regressions, whilst obviously attractive, fares poorly when trying to account for such complex dynamics in the lag structure (as illustrated by the empirically evaluated autocovariance and Cross-autocovariance functions see subsection 2.3.7 in page 47). Hence, we specify a realized estimator in the tradition of Newey-West augmented by the recent results in [Barndorff-Nielsen, Hansen, Lunde, and Shephard, 2008, 2009a, 2011] (BNHLS).

The first objective is to provide an unbiased estimator of $Q = T(y_t y_t')$ and hence derive a consistent estimator of Π.

Following Barndorff-Nielsen, Hansen, Lunde, and Shephard [2011], $\tilde{W}$ is a Newey and West demonstrate. let $\tilde{W} = \mathscr{K}(Y)$, where $\mathscr{K}(Y)$ is a kernel operator on the $T \times N$ data matrix Y,

$$\tilde{W} = \mathscr{K}(Y) := \sum_{h=-n}^{n} K(h/H)\Gamma_h,$$

where Γ_h is the h auto-covariance matrix $\mathbb{E}[y_t y_{t-h}] = \sum_{t=h+1}^{t} y_t y_{t-h}'$. Newey and West demonstrate that $q/T^{\frac{1}{4}} \to 0$, then $\tilde{W}_T \xrightarrow{p} W$.

A key feature here is that by construction $W \equiv \lim_{T\to\infty} T.\mathbb{E}[y_t y_t']$ is finite, as the calendar time is fixed to a single day. Following Barndorff-Nielsen and Shephard [2004b] we place the following assumptions on $K(\cdot)$. (i) $K(0) = 1$, $K'(0) = 0$;, (ii) K is at least twice differentiable with continuous derivatives; (iii) define $K_\bullet^{0.0} = \int_0^\infty K(z)^2 dz$, $K_\bullet^{1.1} = \int_0^\infty K'(z)^2 dz$ and $K_\bullet^{2,2} = \int_0^\infty K''(z)^2 dz$ then $K_\bullet^{0,0}, K_\bullet^{1,1}, K_\bullet^{2,2} < \infty$; (iv) $\int_0^\infty K(z)\exp(iz\lambda)dz \geq 0\ \forall \lambda \in \mathbb{R}$. It is worth restating some of the rationale from Barndorff-Nielsen and Shephard [2004b] to

Table 3.1: The Realized Kernels from BNHLS 2008

Kernel function, $k(x)$			$\lvert k''(0)\rvert$	$k_{\bullet}^{0,0}$	$\lvert k''(0)(k_{\bullet}^{0,0})^2\rvert^{1/5}$
Parzen	$k(x) = \begin{cases} 1-6x^2+6x^3 \\ 2(1-x)^3 \\ 0 \end{cases}$	$0 \le x \le \frac{1}{2}$ $\frac{1}{2} \le x \le 1$ $x > 1$	12	0.269	0.97
Quadratic spectral	$k(x) = \frac{3}{x^2}\left(\frac{\sin x}{x} - \cos x\right)$	$x \ge 0$	1/5	$3\pi/5$	0.93
Fejér	$k(x) = \left(\frac{\sin x}{x}\right)^2$	$x \ge 0$	2/3	$\pi/3$	0.94
Tukey-Hanning$_\infty$	$k(x) = \sin^2\left(\frac{\pi}{2}e^{-x}\right)$	$x \ge 0$	$\pi^2/2$	0.52	1.06
BNHLS (2008)	$k(x) = (1+x)e^{-x}$	$x \ge 0$	1	5/4	1.09

For our Kernels we use the same set as proposed in BNHLS 2008 [a] and this table is adapted from Table 1 of BNHLS 2008. Note that $\lvert k''(0)(k_{\bullet}^{0,0})^2\rvert^{1/5}$ measures the relative asymptotic efficiency of $k \in k$.

[a] Barndorff,Hansen,Lunde,Shephard,2008.

illustrate the importance of these assumptions in maintaining the properties of $R(z)$. For assumption (i) when, $K(0) = 1$ results in Γ_0 having unit weight in the estimator, while $K'(0) = 0$ means the kernel gives close to unit weight to Γ_h for small values of $|h|$. Assumption (iv) guarantees $\mathscr{K}(Y)$ to be at least positive semi-definite. For our purposes we require $\mathscr{K}(Y)$ to be positive definite. We will review the reduced rank case in chapter 4. We will show that the standard OLS estimator radically under estimates the size of the effect (by an order of magnitude) and the estimates are robust to choice of kernel and are verifiable by simulation.

In 3.1 we demonstrate how the Kernels are illustrated under the Newey-West, as shown in [Barndorff-Nielsen, Hansen, Lunde, and Shephard, 2008].

3.1.1 Spectral-based estimators

We can see that whilst elements of y_t are highly persistent prima facie evidence is that the columns of Y are covariance-stationary which been shown in

[Barndorff-Nielsen, Hansen, Lunde, and Shephard, 2008] . Let the theoretical autocovariance-generating function $G_Y(z) = \Sigma_{h=-\infty}^{\infty}\Gamma_h z^h$ be evaluated at $z = 1$, or, equivalently as 2π times the population spectrum at frequency zero:

$$\tilde{W} = \sum_{h=-\infty}^{\infty} \Gamma_h = 2\pi w_Y(0)$$

We then estimate a r-order vector autoregression and compute the residual auto covariance matrices $\hat{\Gamma}_t^u$, where $\hat{\Gamma}_t^u$ is presumed to be the residual autocorrelation not entirely captured by the VAR. The second step in the procedure is to estimate $\tilde{W}^*$ using via $\mathscr{K}(Y)$. Thus:

$$\tilde{W}^* = \sum_{h=-n}^{n} k\left(\frac{h}{H}\right)\Gamma_h,kar$$

where

$$\sum_{j=h+1}^{n} x_j x'_{j-h,} \quad h \geq 0$$

and where h is a parameter representing the maximal order of autocorrelation assumed for v_t. The matrix $\tilde{W}_T^*$ maybe decomposed by $2\pi.w_v(0)$, where $s_v(\omega)$ is the spectral density of v:

$$w_v(\omega) = (2\pi)^{-1} \sum_{v=-\infty}^{\infty} \{E(v_t v'_{t-v})\}e^{-i\omega v}$$

The original series y_t can be obtained from v_t by applying the following filter:

$$y_t = [I_n - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p]^{-1} v_t$$

The spectral density of y is related to spectral density of v according to

$$w_y(\omega) = \{[I_n - \Phi_1 e^{-i\omega} - \Phi_2 e^{-2i\omega} - \cdots - \Phi_p e^{-pi\omega}]\}^{-1} w_v(\omega) \tag{3.1}$$

$$\times \{[I_n - \Phi_1 e^{i\omega} - \Phi_2 e^{2i\omega} - \cdots - \Phi_p e^{pi\omega}]'\}^{-1} \tag{3.2}$$

Hence, an estimate of 2π times the spectral density of y at frequency zero is given by

$$\tilde{W}_T = \{[I_n - \hat{\Phi}_1 - \hat{\Phi}_2 - \cdots - \hat{\Phi}_p]\}^{-1} \hat{W}_T^*$$
$$\times \{[I_n - \hat{\Phi}_1 - \hat{\Phi}_2 - \cdots - \hat{\Phi}_p]'\}^{-1}$$

where $\tilde{S}_T^*$ is calculated from the Kernel density estimator.

3.1.2 Adjusting the Standard Newey-West Approach

Let $\bar{y} = (1/T) \sum_{t=1}^{T} y_t$ be the vector of sample means for elements of the vector process $y_t \in \mathbb{R}^n$. Recall that $\mathbf{F}$ is the 'true' companion matrix for the VAR process describing the evolution of y_t and the matrices $\Pi_1, \ldots, \Pi_p$ the lag operator notation of the process is given by $C(L)y_t = c + u_t$ and the 'true' mean of the process is given by the coefficients $\Pi_1, \ldots, \Pi_r$ and the constant c.

We can now write down the moments for the covariance stationary VAR model in terms of the autocovariance functions and then extend over a variety of moments to recover the asymptotic properties of the estimator under a variety of

conditions. Defining the following notation :

$$N\mathbb{E}[\bar{y} - \mu] = \mathbf{0} \tag{3.3}$$

$$N\mathbb{E}[(\bar{y} - \mu)(\bar{y} - \mu)'] = \mathbf{S}^* \tag{3.4}$$

$$N\mathbb{E}[(\bar{y} - \mu)((\bar{y} - \mu) \otimes (\bar{y} - \mu))'] = \mathbf{Q}^* \tag{3.5}$$

$$N\mathbb{E}[(\bar{y} - \mu)((\bar{y} - \mu) \otimes (\bar{y} - \mu) \otimes (\bar{y} - \mu))'] = \mathbf{P}^* \tag{3.6}$$

where $\otimes$ is the Kronecker product. Newey-West use a simple linear Kernel to generate a HAC [1], when the estimate is $\mathbf{S}^*$, this is given by :

$$(1/N)\mathbf{S}^* = \sum_{h=1}^{H} \left(1 - \frac{h}{H+1}\right) (\hat{\Gamma}_h + \hat{\Gamma}'_h) \tag{3.7}$$

To extending this to a more general framework, we propose the following equations for the stochastic structure of the VAR:

$$\mu = (\mathbf{I} - \sum_{h=1}^{r} \Pi_h)^{-1} c \tag{3.8}$$

$$(1/N)\hat{\mathbf{S}}^* =^p k(0)\hat{\Gamma}_0 + \sum_{h=1}^{H} k(h/H)(\hat{\Gamma}_h) + k(h/H)(\hat{\Gamma}'_h) \tag{3.9}$$

$$(1/N)\hat{\mathbf{Q}}^* =^p k(0)\hat{\Xi}_0 + \sum_{h=1}^{H} k(h/H)(\hat{\Xi}_h) + k(h/H)(\hat{\Xi}_{-h}) \tag{3.10}$$

$$(1/N)\hat{\mathbf{P}}^* =^p k(0)\hat{\Lambda}_0 + \sum_{h=1}^{H} k(h/H)(\hat{\Lambda}_{-h}) + k(h/H)(\hat{\Lambda}_{-h}), \tag{3.11}$$

[1] HAC estimate is use for OLS time-series with autocorrelations and hetroscedasticity

with the autoskewness and autokurtosis matrices given by:

$$\Xi_h = (1/N)\sum_{t=h}^{T} y_t(y_{t-h} \otimes y_{t-h}), \quad \Xi_h = (1/N)\sum_{t=h}^{T} y_t(y_{t-h} \otimes y_{t-h} \otimes y_{t-h}) \tag{3.12}$$

Finally, we compute the model moments (Adjusted Newey-West) by:

$$\mu = (\mathbf{I} - \sum_{h=1}^{r} \Pi_h)^{-1} c \tag{3.13}$$

$$(1/N)\hat{\mathbf{S}}^* =^p k(0)\hat{\Gamma}_0 + \sum_{h=1}^{H} k(h/H)(\hat{\Gamma}_h) + k(h/H)(\hat{\Gamma}'_h) \tag{3.14}$$

$$(1/N)\hat{\mathbf{Q}}^* =^p k(0)\hat{\Xi}_0 + \sum_{h=1}^{H} k(h/H)(\hat{\Xi}_h) + k(h/H)(\hat{\Xi}_{-h}) \tag{3.15}$$

$$(1/N)\hat{\mathbf{P}}^* =^p k(0)\hat{\Lambda}_0 + \sum_{h=1}^{H} k(h/H)(\hat{\Lambda}_{-h}) + k(h/H)(\hat{\Lambda}_{-h}), \tag{3.16}$$

which will are now matched to the moments from the actual model:

$$\bar{y} = (1/T)\sum_{t=r+1}^{T} y_t \tag{3.17}$$

$$\hat{\mathbf{S}}^\dagger = \sum_{t=r+1}^{T} (\Pi x_t)(\Pi x_t)' - N\bar{y}\bar{y}' \tag{3.18}$$

$$\hat{\mathbf{Q}}^\dagger = \sum_{t=r+1}^{T} (\Pi x_t)((\Pi x_t) \otimes (\Pi x_t))' - N\bar{y}(\bar{y} \otimes \bar{y})' \tag{3.19}$$

$$\hat{\mathbf{P}}^\dagger = \sum_{t=r+1}^{T} (\Pi x_t)((\Pi x_t) \otimes (\Pi x_t) \otimes (\Pi x_t))' - N\bar{y}(\bar{y} \otimes \bar{y} \otimes \bar{y})' \tag{3.20}$$

Notice that the terms in $\hat{\mathbf{S}}^*$ are demeaned by $\bar{y}$,

3.1.3 Monte-Carlo simulation example

In this section I will show how the estimator recovers the parameters when the underlying uncertainty process exhibits substantial auto and cross co-skewness and co-kurtosis. The object is to detect the degree of bias and the degree of correction inherent within the model structure.

Consider the bivariate example, setting $r = 1$ and $c = [1/2, 1/2]'$ the generating coefficient Π_0 to be:

$$\Pi_0 = \begin{bmatrix} -\frac{5}{3\pi} & -\frac{4}{3\pi} \\ \frac{4}{3\pi} & \frac{5}{3\pi} \end{bmatrix} \tag{3.21}$$

which has eigenvalues of $\{-1/\pi, 1/\pi\}$ and hence is stationary. To generate autocovariance, co-skewness and co-kurtosis we generate the residuals using the following two-step method. First we generate a set of IID disturbances from a Skewed normal distribution $z_{i,t} \sim \mathcal{SN}(\theta_\zeta, \theta_\alpha, \theta_\epsilon)$, where $\theta_\zeta, \theta_\alpha$ and θ_ϵ are the location, scale and shape parameters respectively. These are then collected into the vectors $z_t = [z_{1,t}, \ldots, z_{n,t}]'$. We generate a set of spectral moving average matrices as follows:

$$\Psi_h = \begin{bmatrix} \exp(-(h/H^\ddagger))\cos(h/H^\ddagger) & -\exp(-(h/H^\ddagger))\sin(h/H^\ddagger) \\ \exp(-(h/H^\ddagger))\sin(h/H^\ddagger) & \exp(-(h/H^\ddagger))\cos(h/H^\ddagger) \end{bmatrix} \tag{3.22}$$

and compute $u_t^* = \sum_{h=0}^{H} \Psi_h z_{t-h}$. We will vary H^* to generate different levels of contamination within the autocorrelation structure. Our simulation runs using the following steps.

1. Choose an integer value of H^* from Ψ_h

2. Pre generate 5,000 sets of draws of the 2-length vector u_t^*, each of $6,000$ observations in length.

3. Pre-compute the MA(1) to MA(2000) matrices from Π^{*1} to Π^{*2000}.

4. Generate a draw of $y_t^* = u_t^* + \sum_{h=1}^{t-1} \Pi^{*h} u_{t-h}^*$ and store in a matrix $Y^* = [y_t']_{t=1}^{6000}$.

5. Delete the first 1,000 rows of Y^*.

6. Estimate $\hat{\Pi}^*$ for each of the 5,000 sets using either OLS or the spectral method above with a specific kernel and compute the quantity $\varepsilon^* = ||\hat{\Pi}^* - \Pi_0||$ and recover the median, 2.5 and 97.5 percentile.

7. Plot the recovered quantiles against the chosen $H^\ddagger$.

When the shape parameter $\theta_\alpha = 0$ and $H^\ddagger = 0$, the simulation adheres to the standard OLS limit theorems. Hence we run the simulation over a selection of $\theta_\alpha = 0$ for comparison. The results of this exercise are given in Figure 3.1.

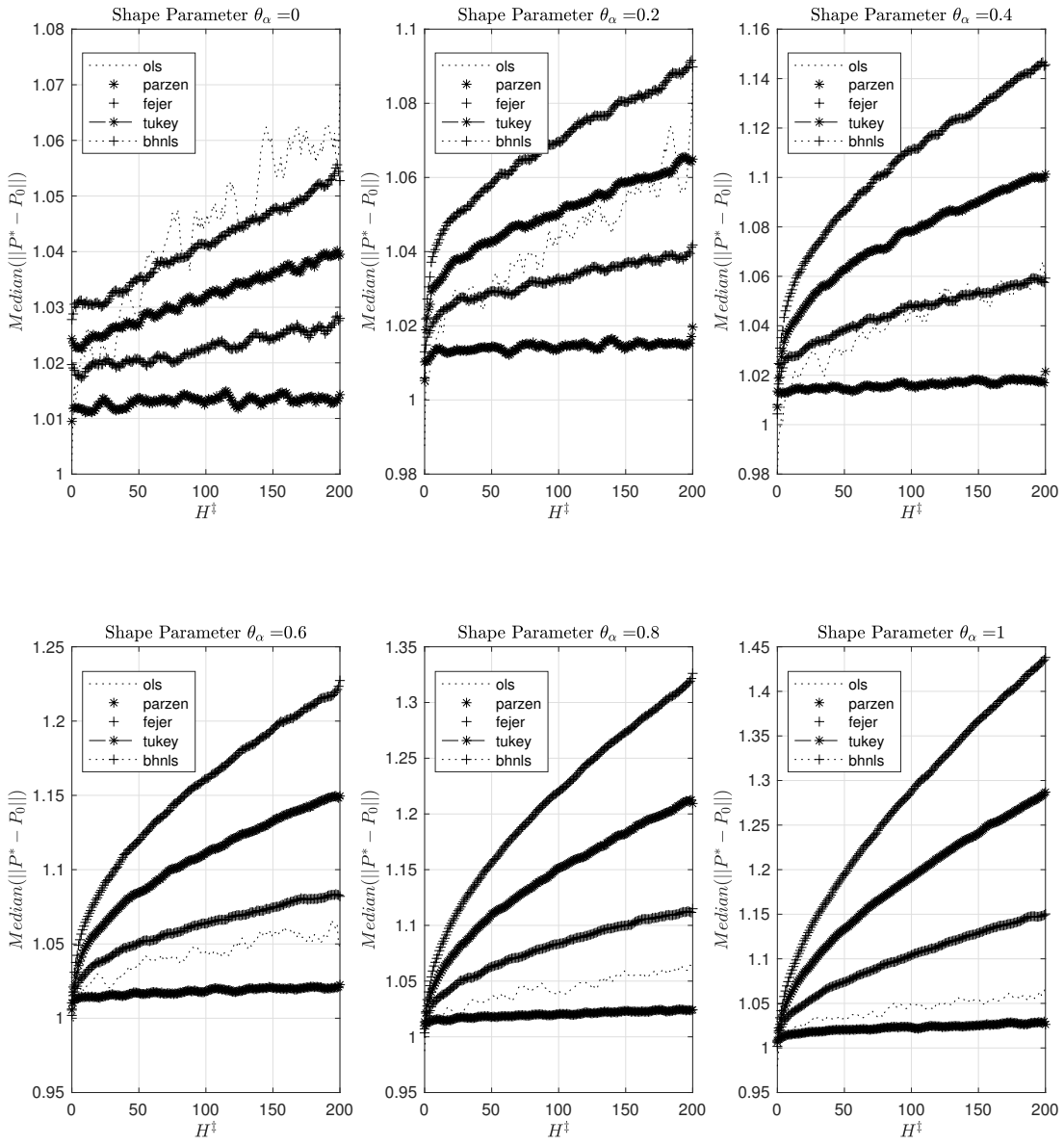


Figure 3.1: Results of the simulation exercise for the new estimator in section 3,1,2

Notes: The individual plots mark each individual choice of θ_α ranging from 0 to 1, in increments of 1/5.

we learned so far from Fig 3.1 for the Monte-Carlo simulation. The six sets of plot-lines on each represents the OLS estimate and the five Kernels we have chosen. The bandwidth of the Kernels is set to $N^{3/5}$ in each case. The abscissa values mark increasing degree of auto- and cross auto- covariance between the variables in the system of equations. The ordinate axis reports the median (unbroken line) and the 2.5 and 97.5 percentiles (thin dotted lines) of the $p = 2$ norm $\varepsilon^* = ||\hat{\Pi}^* - \Pi_0||$ between the estimated and true VAR coefficients.

3.1.4 A Bootstrapping procedure for the Impulse Responses and Block exogeniety tests

Let us now consider a finite sample Monte-Carlo analysis of the estimators consistency. We choose to look at the first two levels $l = 2$ of the order- book and 10 lags in update time. The choice of lag structure is somewhat problematic as the standard measures of over-parameterization may or may not provide an accurate picture. We follow the bootstrap in-bootstrap approach of Killian [1998]; Killian and Demiroglu [2000] and design a bootstrapping procedure to construct an estimate of any bias of the estimates of $\hat{\Pi}$ and to compute the sample confidence intervals for the resulting impulse response functions.

The bootstrapping procedure is as follows. First we estimate $\hat{\Pi}$ directly from the data using the various Kernels mentioned earlier in this chapter. Our benchmark case is when $H = 0$, the estimate of $\hat{\Pi}$ is the standard OLS estimator as shown earlier in this chapter. Following Barndorff-Nielsen and Shephard [2004b] and Voev and Lunde [2007] for the rest of the Kernels we set $H = T^{3/2}$. For our purposes this is somewhat arbitrary as we wish to be less specific about the data- generating process (DGP). Our basic requirement for the DGP is that the unknown 'true' covariance matrix Σ is positive semi-definite and the disturbances U are matrix multivariate normal with at least finite eighth-order moments. We check that the roots of the companion matrix F lie inside the unit circle, which for our type of data are uniformly the case. If the roots of F lie outside the unit circle then the estimates of $\hat{\Pi}$ will of course be super-consistent, and at this point the bias correction is unnecessary. However the model may potentially require an error correction adjustment.

Killian [1998] proposes a bootstrap-in-bootstrap approach to (a) correcting for bias in the parameter estimates and subsequently (b) adjusting for non-normality in the standard errors of the resulting impulse responses. Following Killian [1998] and Killian and Demiroglu [2000] it is useful to set $\pi = \text{vec}(\Pi)$ and $\hat{\pi} = \text{vec}(\hat{\Pi})$ as the vector of parameter estimates from the initial vector autoregression. We then set $\Psi_s(\hat{\pi})$ to be the resulting impulse response at the s forward step. Once we have the initial parameter estimates $\hat{\pi}$ we proceed with the following steps.

Step 1: Generate the bootstraps

Draw N_B draws of the disturbances, denoted U^*, of length $2T$ with the following bootstrap devices:

1. Baseline: $U^* \sim \mathcal{N}(\mathbf{0}_{T\times n}, I_T \otimes \Sigma)$. Disturbance terms are normally distributed with no heteroskedasticity or autocorrelation.

2. Stochastic covariation: $U^* \sim \mathcal{N}(\mathbf{0}_{T\times n}, \tilde{\Omega})$, where the matrix $\tilde{\Omega}$ has block diagonals of $\Omega_{ii} = \tilde{\Sigma}_i$ and each covariance matrix is drawn from a zero centred Wishart distribution, $\tilde{\Sigma}_i \sim \mathcal{W}(q, p, \Sigma)$. For the degrees of freedom q and p we set a relatively high level of variation in the block diagonals by setting $q = p = n$.

3. Stochastic auto-covariation: $U^* \sim \mathcal{N}(\mathbf{0}_{T\times n}, \tilde{\Omega})$ where $\tilde{\Omega}$ is a single draw from a Wishart distribution, such that $\tilde{\Omega} \sim \mathcal{W}(q, p, I_{n\times n} \otimes \Sigma)$ and $p = q = n$. Notice that $\tilde{\Omega}$ is by construction rank deficient. However the covariance $\mathbb{E}(u_i u_i')$, will be generated by a full rank covariance matrix. This provides a very challenging type of contamination for our estimator series of estimators.

4. $U^* \sim boot$, standard bootstrap. The rows $[\hat{u}'_t]_{t=1}^T$ are resampled with replacement to generate $[u_t^*]$.

5. $U^* \sim wboot$, correlation preserving wild-bootstrap. Setting $u_t = [u_{i,t}]_{i=1}^n$ the residuals are resampled by $u_t^* = \hat{u}_t w_t$ where $w_t \sim \mathcal{N}(0,1)$ is an independent draw from a standard normal distribution.

Next generate the simulated draws of $Y^* := [y_t^*]_{t=1}^T$ from $U^* := [u_t^*]_{t=1}^T$ as follows, for the initial r observations $y^*_{t\in\{1,\dots,r\}} = u_t^* + \sum_{s=1}^t \Psi(\hat{\pi})_s u^*_{t-s}$ and for the remaining $2T-r$ observations we compute $y^*_{t\in\{t>r,\dots,2T\}} = u_t^* + \sum_{i=1}^r \Pi_i y_{t-i}$. We then discard the initial $T-r$ observations and construct the lagged matrix X^*. For each draw then compute:

$$\hat{\Pi}^* = R^{-1}(x^*)R(x^*,y^*), \quad and \quad \hat{\pi}^* = \text{vec}(\hat{\Pi}^*)$$

and approximate the bias $\phi = \mathbb{E}(\hat{\pi} - \pi)$. This is estimated via $\tilde{\phi} \approx \mathbb{E}(\hat{\pi}^* - \hat{\pi}) = N_B^{-1}\sum_{j=1}^{N_B}\hat{\pi}_j^* - \hat{\pi}$. The bias-corrected coefficients are denoted $\tilde{\pi} = \hat{\pi} - \tilde{\phi}$. To check the stationarity of the bias-corrected coefficients we compute the companion matrix $\tilde{F}$ and compute the roots. If any of the roots lie outside the unit circle we follow Killian [1998] and rescale by the following iterative procedure $\tilde{\pi}_{k=1} = \hat{\pi} - \hat{\phi}$, subsequently $\tilde{\pi}_{k+1} = \hat{\pi} - \delta_{k+1}\hat{\phi}$, where $\delta_1 = 1$ and $\delta_{k+1} = \delta_k - \varepsilon$. Killian [1998] sets $\varepsilon = 0.01$. However we find that 10e-5 provides a good trade-off in terms of ensuring that persistent processes with near unit roots are appropriately captured and the speed of correction.[1]

[1] From 2,840 days of data the root correction needed to only be implemented twice.

3.2 Demonstration results

We apply the spectral least squares estimator to our Oil Futures data for the set of days in our sample in chapter 2, outlined in Table 2.1. We utilize the Schwartz Bayesian Criterion to assess the appropriate number of lags from the raw residuals. In most cases the appropriate lag structure consists of over a dozen lags. we plot the entire set of impulse responses, but, our main interest lies specifically with the mid-price, update speed and the bid-ask spread, as these are the critical ingredients for formulating a HFT strategy.

For brevity in Tables 3.2 to 3.6 we present only the first-order autoregressive matrix $\hat{\Pi}_1$ to illustrate that the kernels generate slightly different results from the estimation procedure. To generate the standard errors denoted std.err$(\hat{\pi}_{i,j})$, we compute the Hessian matrix from the minimization procedure and invert to compute the Fisher information matrix (J-matrix in GMM terms) and then extract the square-root of the diagonal terms. As the Hessian is computed directly during the optimization this is a computationally simple approach. In certain cases the Hessian needs to be reconditioned and we utilize a step size of the squareroot of 2×10^{-16}, which is the squareroot of the smallest 64bit floating point number to numerically recover the Hessian.

For the number of levels in the order-book, we choose the first three (best price, next best price and second next best price) for computing the spreads and the order-flow imbalance (proxying relative depth) as these appear to be the most actively quoted. Hence $L = 3$; experimentation with $L = 4$ does not yield materially different results.

We have a larger set of results from days not included in Table 2.1. The

objective of the sampling is to provide a consistent set of estimates for the active contracts for a given day. As such the term structure has a degree of activity equivalent to the sample utilized herein and variation is consistent across the greater sample of days covered. Hence we would suggest that the sample used is representative of the overall market.

All results below are sampled from the days outlined in Table 2.1 and the standard errors are computed by resampling using the bootstrap routine outlined in in the previous subsection.

3.2.1 Impulse response analysis

To present the impulse response plots, we do not plot the steps, but against the update speed (τ_t) measured in milliseconds, hence the adjustment in milliseconds is always on the abscissa values as this helps put context on the adjustment speed.

Figures 3.2 to 3.8 present the impulse responses for the six (OLS, Parzen, Qspec, Fajer,Tukey,Bhnls) kernels in our analysis and the OLS estimates. Note that whilst the OLS estimates appear to be a straight line, they are not. Without the kernel adjustments the OLS estimates are biased towards zero, following the logic of the Epps effect[1]. Hence, the IRFs are about one to two orders of magnitude smaller in terms of extracting the autoregressive structure. Indeed, this is visibly the case from the auto-and cross auto-covariance plots as the level of lagged dependency, the rate of decay and the complexity are respectively high (far greater than 50%) and, slow (over 1,000 lags) and complex (lots of oscillating

[1] Epps effect is the phenomenon that the empirical correlation between the returns of two different stocks decreases with the length of the interval for which the price changes are measured. The phenomenon is caused by non-synchronous/asynchronous trading and discretization effects.

Table 3.2: First Order Autoregressive Matrix, Spectral Least Squares Using a Parzen Kernel.

	$\pi_{j,1}$	$\pi_{j,2}$	$\pi_{j,3}$	$\pi_{j,4}$	$\pi_{j,5}$	$\pi_{j,6}$
$\pi_{1,i}$	0.02973***	0.00070	0.02173***	0.02177***	-0.02426***	-0.01986***
std.err($\pi_{1,i}$)	(0.00071)	(0.00070)	(0.00070)	(0.00070)	(0.00070)	(0.00070)
$\pi_{2,i}$	0.00002	0.01069***	0.00019	0.00013	-0.00006	-0.00004
std.err($\pi_{2,i}$)	(0.00012)	(0.00012)	(0.00012)	(0.00012)	(0.00012)	(0.00012)
$\pi_{3,i}$	-0.00037	-0.00041	0.01035***	-0.00018	-0.00001	-0.00003
std.err($\pi_{3,i}$)	(0.00038)	(0.00038)	(0.00038)	(0.00038)	(0.00038)	(0.00038)
$\pi_{4,i}$	-0.00019	0.00029	-0.00007	0.01041***	0.00018	0.00017
std.err($\pi_{4,i}$)	(0.00037)	(0.00037)	(0.00037)	(0.00037)	(0.00037)	(0.00037)
$\pi_{5,i}$	0.00049***	0.00000	0.00000	0.00001	0.01024***	-0.00002
std.err($\pi_{5,i}$)	(0.00011)	(0.00011)	(0.00011)	(0.00011)	(0.00011)	(0.00011)
$\pi_{6,i}$	0.00025**	-0.00001	0.00002	0.00002	-0.00014	0.01011***
std.err($\pi_{6,i}$)	(0.00010)	(0.00010)	(0.00010)	(0.00010)	(0.00010)	(0.00010)

Notes This presents the 6×6 estimates for the first order autoregressive matrix $\hat{\Pi}_1 = [\hat{\pi}_{i,j}]$ estimated by spectral least squares using a Parzen kernel with corresponding standing errors. The asterisks $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1%. The kernel uses a bandwidth parameter of $N^{3/5}$. Refer to Table 2.1 for sample characteristics.

Table 3.3: First Order Autoregressive Matrix, Spectral Least Squares Using a Quadratic Spectral (qspec) Kernel.

	$\pi_{j,1}$	$\pi_{j,2}$	$\pi_{j,3}$	$\pi_{j,4}$	$\pi_{j,5}$	$\pi_{j,6}$
$\pi_{1,i}$	0.03108***	0.00082	0.02478***	0.02484***	-0.02786***	-0.02295***
std.err($\pi_{1,i}$)	(0.00073)	(0.00073)	(0.00073)	(0.00073)	(0.00073)	(0.00073)
$\pi_{2,i}$	0.00003	0.01069***	0.00020	0.00014	-0.00007	-0.00004
std.err($\pi_{2,i}$)	(0.00012)	(0.00011)	(0.00011)	(0.00011)	(0.00011)	(0.00011)
$\pi_{3,i}$	-0.00038*	-0.00044	0.01033***	-0.00020	-0.00001	-0.00003
std.err($\pi_{3,i}$)	(0.00037)	(0.00037)	(0.00037)	(0.00037)	(0.00037)	(0.00037)
$\pi_{4,i}$	-0.00019	0.00031	-0.00007	0.01041***	0.00019	0.00018
std.err($\pi_{4,i}$)	(0.00036)	(0.00036)	(0.00036)	(0.00036)	(0.00036)	(0.00036)
$\pi_{5,i}$	0.00050***	0.00000	0.00001	0.00002	0.01022***	-0.00003
std.err($\pi_{5,i}$)	(0.00010)	(0.00010)	(0.00010)	(0.00010)	0.00010)	(0.00010)
$\pi_{6,i}$	0.00026**	-0.00001	0.00003	0.00003	-0.00015	0.01010***
std.err($\pi_{6,i}$)	(0.00010)	(0.00010)	(0.00010)	(0.00010)	0.00010)	(0.00010)

Notes This presents the 6×6 estimates for the first order autoregressive matrix $\hat{\Pi}_1 = [\hat{\pi}_{i,j}]$ estimated by spectral least squares using a Quadratic Spectral (qspec) Kernel with corresponding standing errors. The asterisks $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1%. The kernel uses a bandwidth parameter of $N^{3/5}$. Refer to Table 2.1 for sample characteristics.

Table 3.4: First Order Autoregressive Matrix, Spectral Least Squares Using a Féjer Kernel.

	$\pi_{j,1}$	$\pi_{j,2}$	$\pi_{j,3}$	$\pi_{j,4}$	$\pi_{j,5}$	$\pi_{j,6}$
$\pi_{1,i}$	0.02315***	0.00014	0.00654**	0.00652***	-0.00674**	-0.00511***
std.err($\pi_{1,i}$)	(0.00059)	0.(00058)	(0.00058)	(0.00058)	(0.00058)	(0.00058)
$\pi_{2,i}$	0.00002	0.01069***	(0.00007	0.00004	-0.00002	-0.00001
std.err($\pi_{2,i}$)	(0.00015)	(0.00015)	(0.00015)	(0.00015)	(0.00015)	(0.00015)
$\pi_{3,i}$	-0.00029	-0.00012	0.01058	-0.00001	-0.00002	-0.00004
std.err($\pi_{3,i}$)	(0.00053)	(0.00052)	(0.00052)	(0.00052)	(0.00052)	(0.00052)
$\pi_{4,i}$	-0.00025	0.00007	-0.00021	0.01034***	0.00007	0.00008
std.err($\pi_{4,i}$)	(0.00052)	(0.00052)	(0.00052)	(0.00052)	(0.00052)	(0.00052)
$\pi_{5,i}$	0.00045*	-0.00001	-0.00003***	-0.00002**	0.01031***	0.00001
std.err($\pi_{5,i}$)	(0.00015)	(0.00015)	(0.00015)	(0.00015)	(0.00015)	(0.00015)
$\pi_{6,i}$	0.00020	-0.00001	-0.00001	-0.00001	-0.00007	0.01017
std.err($\pi_{6,i}$)	(0.00015)	(0.00015)	(0.00015)	(0.00015)	(0.00015)	(0.00015)

Notes This presents the 6×6 estimates for the first order autoregressive matrix $\hat{\Pi}_1 = [\hat{\pi}_{i,j}]$ estimated by spectral least squares using a Féjer Kernel with corresponding standing errors. The asterisks $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1%. The kernel uses a bandwidth parameter of $N^{3/5}$. Refer to Table 2.1 for sample characteristics.

Table 3.5: First Order Autoregressive Matrix, Spectral Least Squares Using a Tukey-Hanning Kernel.

	$\pi_{j,1}$	$\pi_{j,2}$	$\pi_{j,3}$	$\pi_{j,4}$	$\pi_{j,5}$	$\pi_{j,6}$
$\pi_{1,i}$	0.01994***	0.00005	0.00216***	0.00216***	-0.00225***	-0.00169***
std.err($\pi_{1,i}$)	(0.00055)	(0.00054)	(0.00054)	(0.00054)	(0.00054)	(0.00054)
$\pi_{2,i}$	0.00002	0.01059***	0.00003	0.00002	-0.00001	0.00000
std.err($\pi_{2,i}$)	(0.00020)	(0.00019)	(0.00019)	(0.00019)	(0.00019)	(0.00019)
$\pi_{3,i}$	-0.00018	-0.00004	0.01059***	0.00002	0.00000	-0.00001
std.err($\pi_{3,i}$)	(0.00068)	(0.00067)	(0.00067)	(0.00067)	(0.00067)	(0.00067)
$\pi_{4,i}$	-0.00023	0.00001	-0.00024	0.01029***	0.00003	0.00004
std.err($\pi_{4,i}$)	(0.00067)	(0.00066)	(0.00066)	(0.00066)	(0.00066)	(0.00066)
$\pi_{5,i}$	0.00033	-0.00001	-0.00003	-0.00002	0.01030***	0.00001
std.err($\pi_{5,i}$)	(0.00020)	(0.00020)	(0.00020)	(0.00020)	(0.00020)	(0.00020)
$\pi_{6,i}$	0.00015	0.00000	-0.00001	-0.00001	-0.00005	0.01016***
std.err($\pi_{6,i}$)	(0.00020)	(0.00019)	(0.00019)	(0.00019)	(0.00019)	(0.00019)

Notes This presents the 6×6 estimates for the first order autoregressive matrix $\hat{\Pi}_1 = [\hat{\pi}_{i,j}]$ estimated by spectral least squares using a Tukey-Hanning Kernel with corresponding standing errors. The asterisks $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1%. The kernel uses a bandwidth parameter of $N^{3/5}$. Refer to Table 2.1 for sample characteristics.

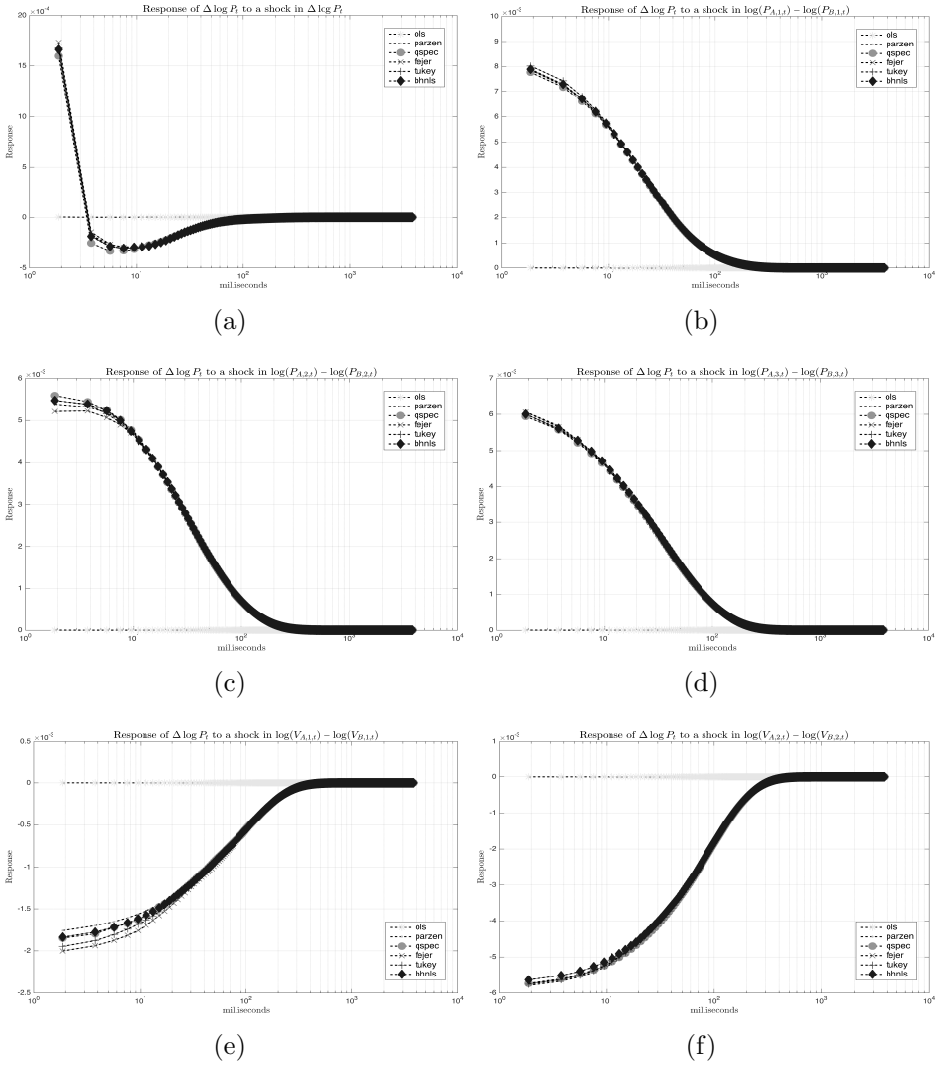


Figure 3.2: Impulse Response Analysis by Kernel Compared to OLS (Returns)
Notes: This represents the Impulse Response Analysis by Kernel Compared to OLS, to the returns, and show the shock in millisecond a) show how the (price) respond to a shock in the price. b)show how the (price) respond to a shock in the price at level 1. c)show how the (price) respond to a shock in the price in level 2. d)show how the (price) respond to a shock in the price in level 3. e)show how the (price) respond to a shock in the volume in level 1. f)show (price) at level 2 respond to a shock in the volume in level 2

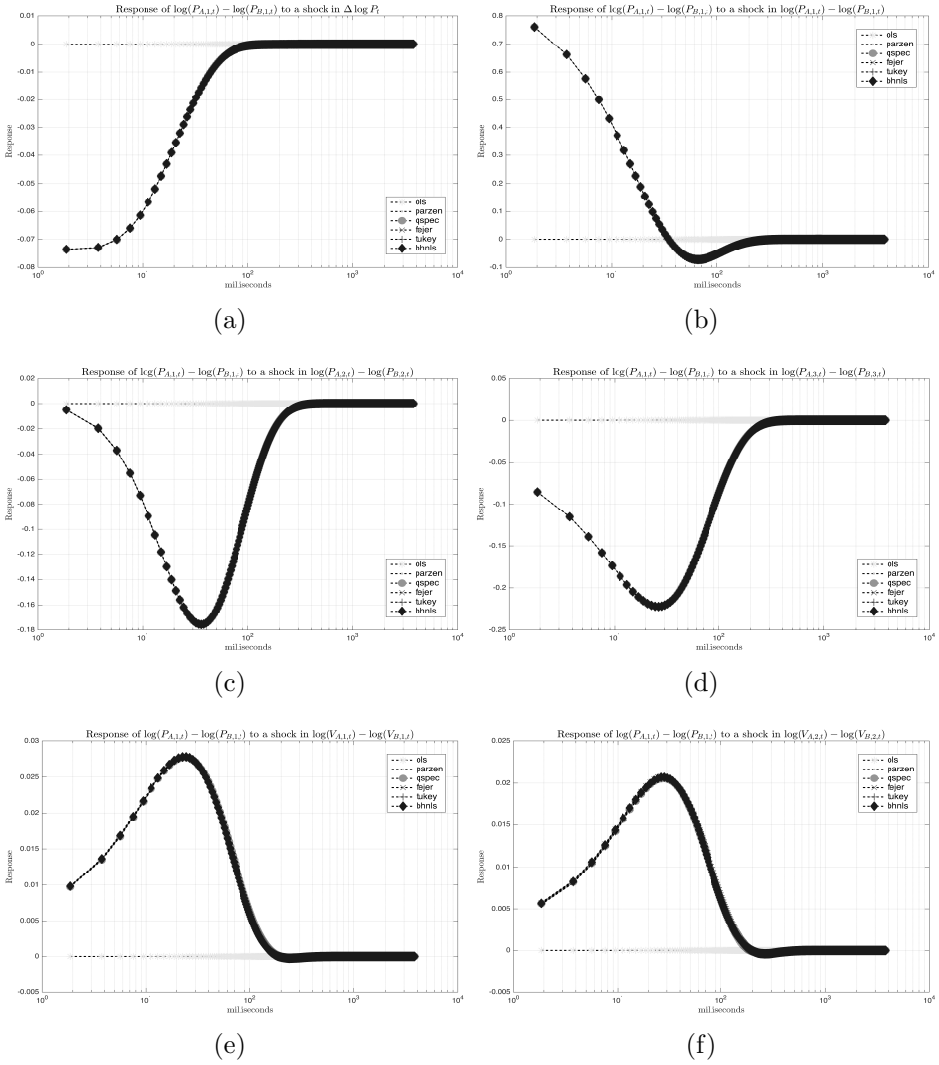


Figure 3.3: Impulse Response Analysis by Kernel Compared to OLS (Change in Asks, and Bids at level 1 Price of the limited order book)

Notes: This represents the Impulse Response Analysis by Kernel Compared to OLS, the impact of the shocks on the level 1 (Ask-Bid) price. a) show how the (Ask, Bid) respond to a shock in the price. b)show how the (Ask, Bid) at level1 respond to a shock in the price in level 1. c)show how the (Ask, Bid) at level1 respond to a shock in the price in level 2. d)show how the (Ask, Bid) at level1 respond to a shock in the price in level 3. e)show how the (Ask, Bid) at level1 respond to a shock in the volume in level 1. f)show how the (Ask, Bid) at level1 respond to a shock in the volume in level 2

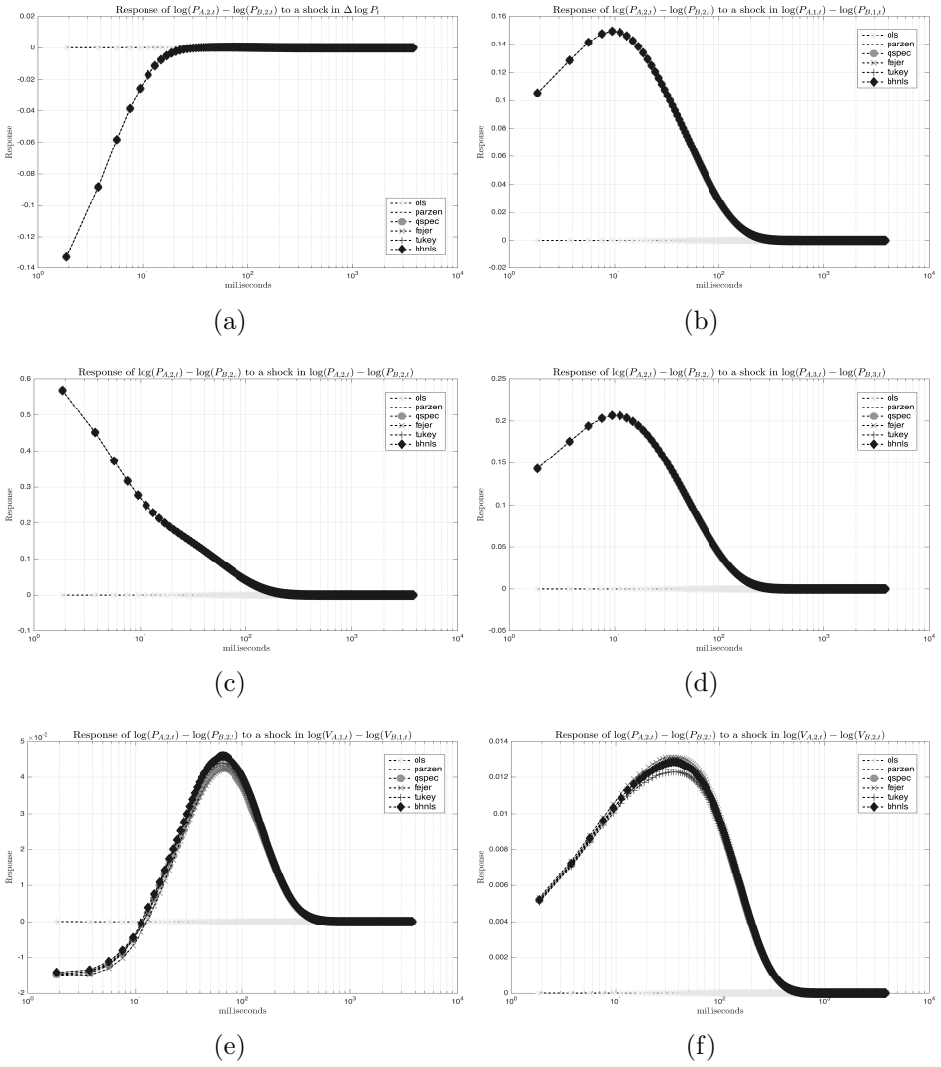


Figure 3.4: Impulse Response Analysis by Kernel Compared to OLS (Change in Asks, and Bids at level2 Price of limited order book)

Notes: This represents the Impulse Response Analysis by Kernel Compared to OLS, the impact of the shocks on the level 2(Ask-Bid) price. a) show how the (Ask, Bid) at level 2 respond to a shock in the price. b)show how the (Ask, Bid) at level 2 respond to a shock in the price at level 1. c)show how the (Ask, Bid) at level 2 respond to a shock in the price in level 2. d)show how the (Ask, Bid) at level 2 respond to a shock in the price in level 3. e)show how the (Ask, Bid) at level 2 respond to a shock in the volume in level 1. f)show how the (Ask, Bid) at level 2 respond to a shock in the volume in level 2

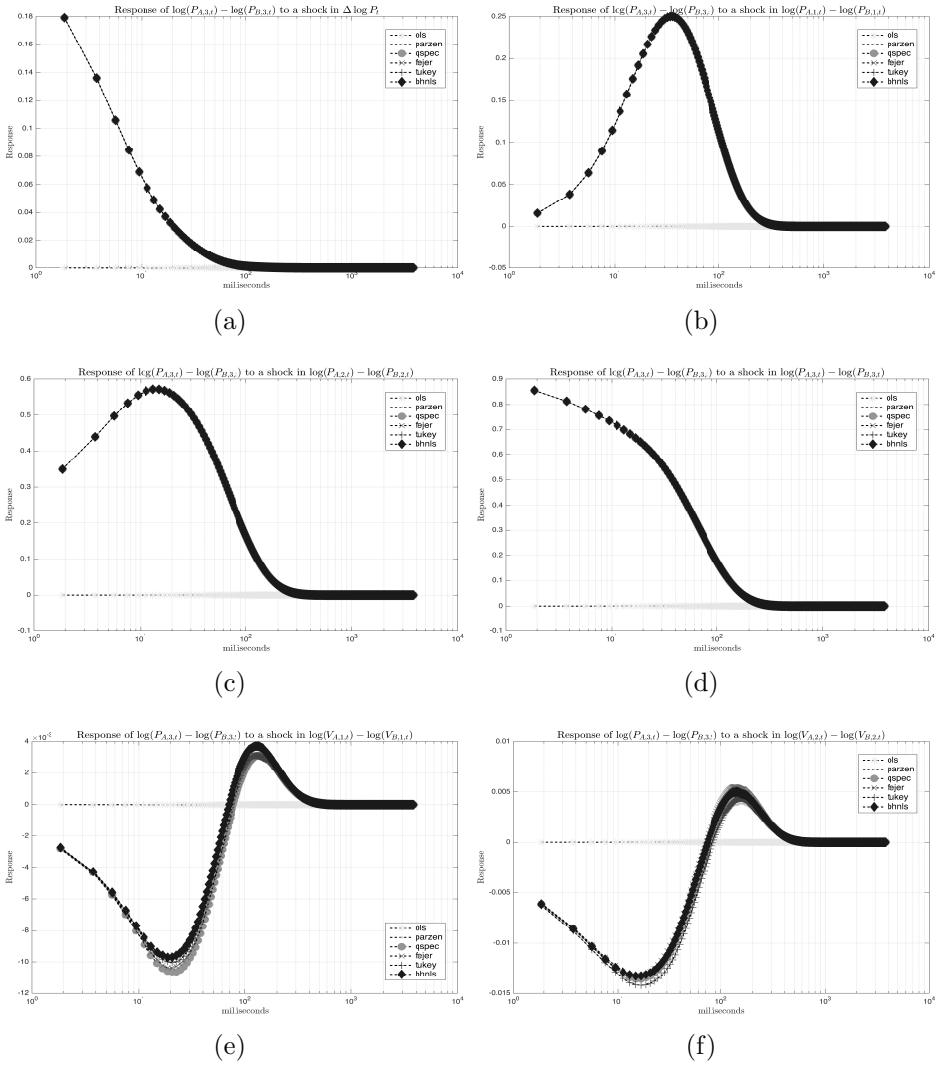


Figure 3.5: Impulse Response Analysis by Kernel Compared to OLS, the impact of the shocks on the level 3(Asks-Bids) Price

Notes: This represents the Impulse Response Analysis by Kernel Compared to OLS, the impact of the shocks on the level 3(Ask-Bid) price. a) show how the (Ask, Bid) at level 3 respond to a shock in the price. b)show how the (Ask, Bid) at level 3 respond to a shock in the price at level 1. c)show how the (Ask, Bid) at level 3 respond to a shock in the price in level 2. d)show how the (Ask, Bid) at level 3 respond to a shock in the price in level 3. e)show how the (Ask, Bid) at level 3 respond to a shock in the volume in level 1. f)show how the (Ask, Bid) at level 3 respond to a shock in the volume in level 2

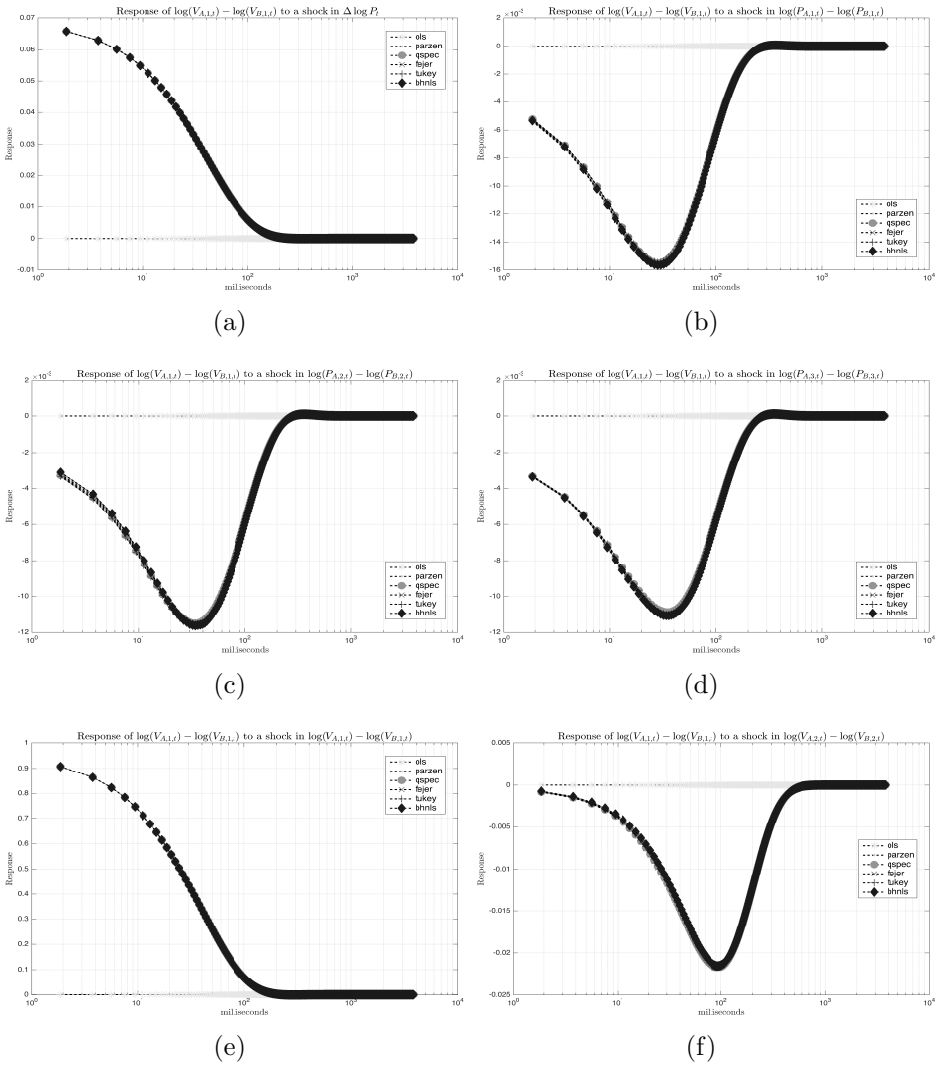


Figure 3.6: Impulse Response Analysis by Kernel Compared to OLS, the impact of the shocks on the level 1 on Volume in balance(Asks-Bids)

Notes: This represents the Impulse Response Analysis by Kernel Compared to OLS, the impact of the shocks on the level 1 on Volume in balance for the (Asks,Bids). a) show how the (Ask, Bid) at level 1 respond to a shock in the price. b)show how the (Ask, Bid) at level 1 respond to a shock in the price at level 1. c)show how the (Ask, Bid) at level 1 respond to a shock in the price in level 2. d)show how the (Ask, Bid) at level 1 respond to a shock in the price in level 3. e)show how the (Ask, Bid) at level 1 respond to a shock in the volume in level 1. f)show how the (Ask, Bid) at level 1 respond to a shock in the volume in level 2

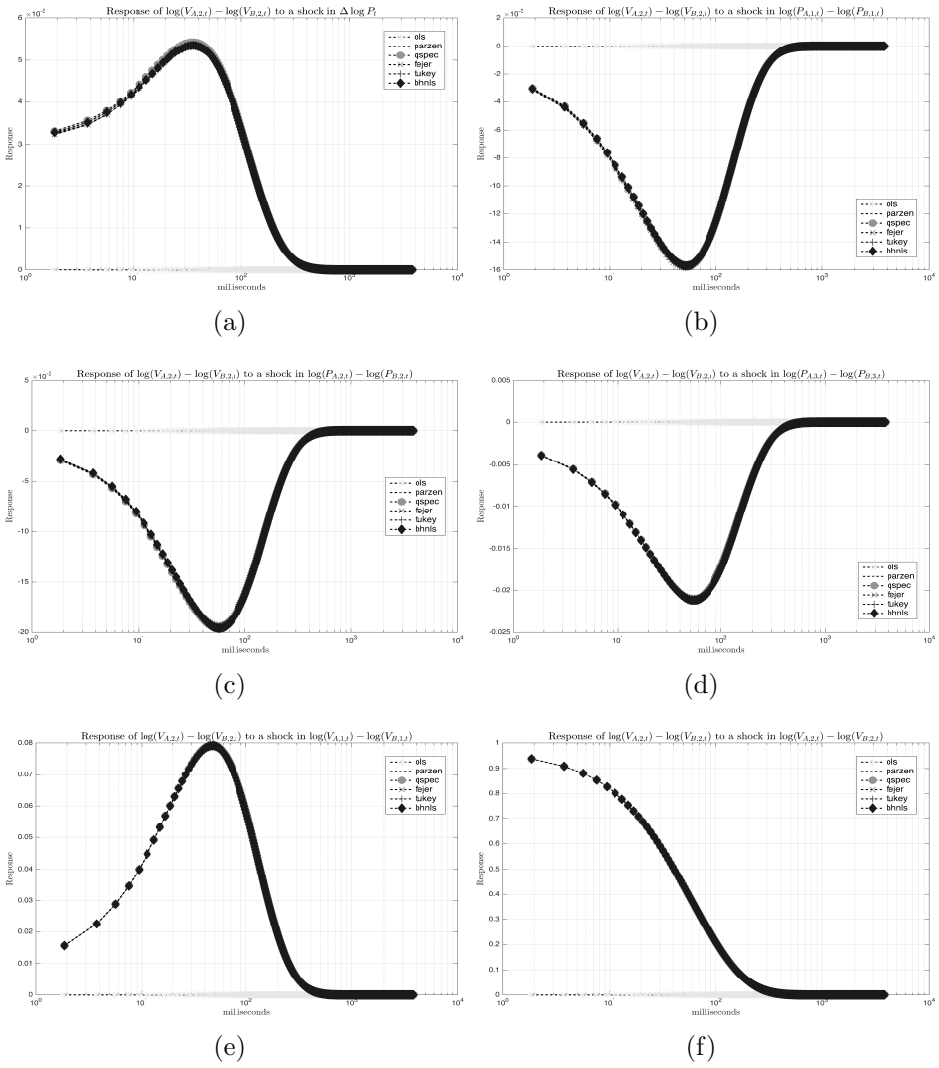


Figure 3.7: Impulse Response Analysis by Kernel Compared to OLS, the impact of the shocks on the level 2 on Volume in balance(Asks-Bids)

Notes: This represents the Impulse Response Analysis by Kernel Compared to OLS, the impact of the shocks on the level 2 on Volume in balance for the (Asks,Bids). a) show how the (Ask, Bid) at level 2 respond to a shock in the price. b)show how the (Ask, Bid) at level 2 respond to a shock in the price at level 1. c)show how the (Ask, Bid) at level 2 respond to a shock in the price in level 2. d)show how the (Ask, Bid) at level 2 respond to a shock in the price in level 3. e)show how the (Ask, Bid) at level 2 respond to a shock in the volume in level 1. f)show how the (Ask, Bid) at level 2 respond to a shock in the volume in level 2

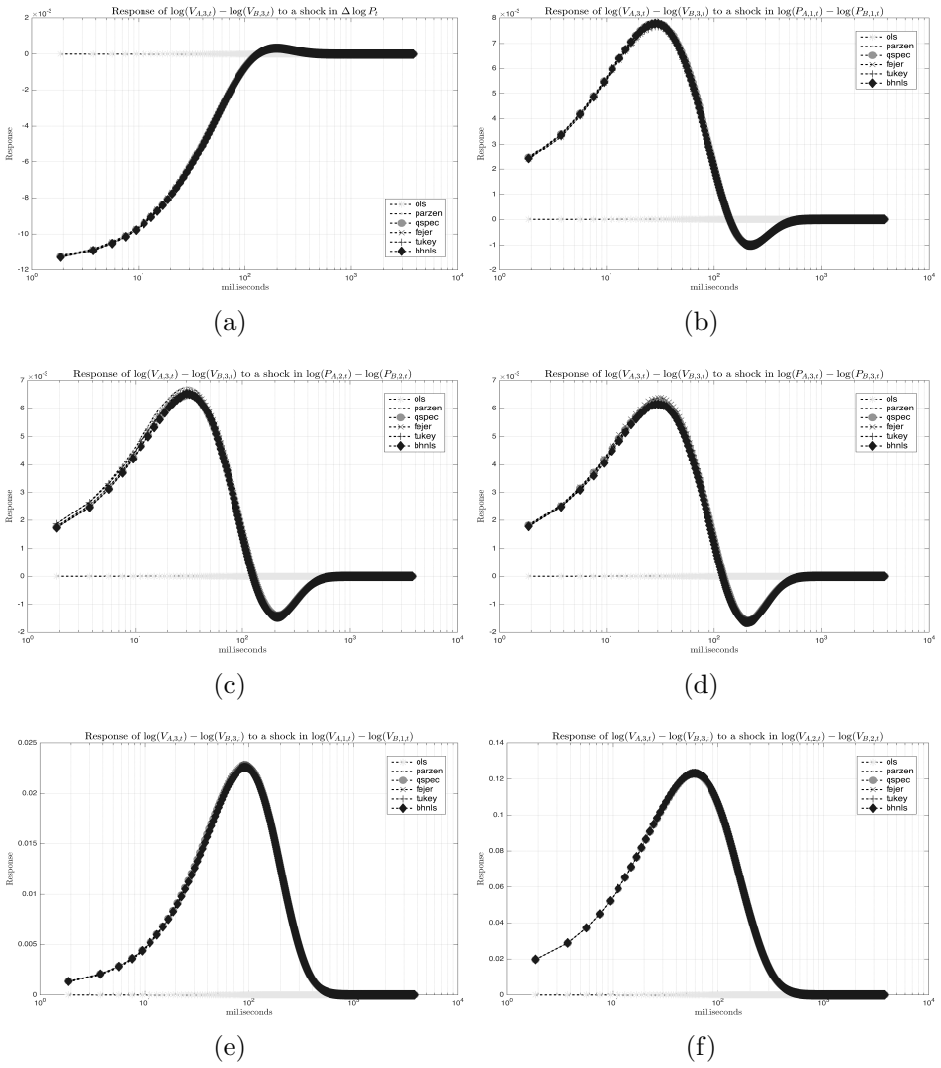


Figure 3.8: Impulse Response Analysis by Kernel Compared to OLS, the impact of the shocks on the level 3 on Volume in balance(Asks-Bids)

Notes: This represents the Impulse Response Analysis by Kernel Compared to OLS, the impact of the shocks on the level 3 on Volume in balance for the (Asks,Bids). a) show how the (Ask, Bid) at level 3 respond to a shock in the price. b)show how the (Ask, Bid) at level 3 respond to a shock in the price at level 1. c)show how the (Ask, Bid) at level 3 respond to a shock in the price in level 2. d)show how the (Ask, Bid) at level 3 respond to a shock in the price in level 3. e)show how the (Ask, Bid) at level 3 respond to a shock in the volume in level 1. f)show how the (Ask, Bid) at level 3 respond to a shock in the volume in level 2

Table 3.6: First Order Autoregressive Matrix, Spectral Least Squares Using a Barndorff-Nielsen, Hansen, Lunde and Shepherd (BNHLS) Kernel.

	$\pi_{j,1}$	$\pi_{j,2}$	$\pi_{j,3}$	$\pi_{j,4}$	$\pi_{j,5}$	$\pi_{j,6}$
$\pi_{1,i}$	0.02412045***	0.00021712	0.00920761***	0.00920507***	-0.0098124***	-0.0076798***
std.err($\pi_{1,i}$)	(0.00061)	(0.00060)	(0.00060)	(0.00060)	(0.00060)	(0.00060))
$\pi_{2,i}$	0.00002	0.01070***	0.00010	0.00007	-0.00003	-0.00001
std.err($\pi_{2,i}$)	(0.00014)	(0.00014)	(0.00014)	(0.00014)	(0.00014)	(0.00014)
$\pi_{3,i}$	-0.00032	-0.00022	0.01049***	-0.00008	-0.00001	-0.00004
std.err($\pi_{3,i}$)	0.00048)	(0.00047)	(0.00048)	(0.00048)	(0.00048)	(0.00048)
$\pi_{4,i}$	-0.00023	0.00014	-0.00014	0.01038***	0.00009	0.00010
std.err($\pi_{4,i}$)	((0.00047)	(0.00047)	(0.00047)	(0.00047)	(0.00047)	(0.00047)
$\pi_{5,i}$	0.00046***	0.00000	-0.00003	-0.00002	0.01030***	0.00001
std.err($\pi_{5,i}$)	(0.00014)	(0.00014)	(0.00014)	(0.00014)	(0.00014)	(0.00014)
$\pi_{6,i}$	0.00022	-0.00001	-0.00001	-0.00001	-0.00009	0.01015***
std.err($\pi_{6,i}$)	0.00013)	(0.00013)	(0.00013)	(0.00013)	(0.00013)	(0.00013)

Notes This presents the 6×6 estimates for the first order autoregressive matrix $\hat{\Pi}_1 = [\hat{\pi}_{i,j}]$ estimated by spectral least squares using a Barndorff-Nielsen, Hansen, Lunde and Shepherd (BNHLS) Kernel with corresponding standing errors. The asterisks $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1%. The kernel uses a bandwidth parameter of $N^{3/5}$. Refer to Table 2.1 for sample characteristics.

auto-covariance and cross auto-covariance).

Whilst the spectral least squares regression generates a similar impulse response structure for each of the kernels in our analysis, there are some discrepancies (this can also be seen in the coefficients for the first-order lagged autoregression matrix, in Tables 3.2 to 3.6, this appears to be in line with the strong signals from the auto-covariance and cross auto-covariance documented previously.

Confidence bounds for each are very tight and difficult to discern in the plots, hence the signal presented are strongly significant. Analysis of the IRFs by sampling suggests that these signals are robust to choice of day and maturity of contract (subject to activity).

Interesting overall features to note are that whilst the signals are strong, the effectively deterministic adjustments in the state of the order book all disappear within 10^2 milliseconds or one tenth of a second. This is interesting when compared to the example in §(2.2) for which the manipulation of the mid-price and spreads by sequentially shocking the order-flow imbalance took place within

900 milliseconds or approximately 1 second. For crude futures any systematic strategy needs to be conducted within about 10 milliseconds to maximise the pay-off as by 10^2 milliseconds the effect of the imbalance shock has died away, see Figures 3.8 subplots (e) and (f). However, for algorithms operating within 10 seconds, the effect is very large excessive supply $V_{A,1,t}$ where (V_A) is the volume of the asks, increasing by one standard deviation affects the mid-price by one order of magnitude more than a shock to the mid-price itself. Cetin, Jarrow, Protter, and Warachka [2006] refer to this as trading at the continuous limit, this is the speed required to effectively front-run the price adjustment from the order-flow imbalance. For crude futures this appears to be around 10 to 10^2 milliseconds. After this any deterministic impact of excessive supply or demand dissappears.

3.2.2 Variation across trading activity and robustness check

As noted previous in Table 2.1 and Figure 2.1, the sample is drawn from a set of days sorted by activity in level 1 of the limit order book, to provide a reasonable number of observations over which to implement the model. Hence some variation may occur when analyzing data from contract-days with less dense activity. We have analyzed a number of additional days, concatenating them together and these are available in an additional almanac of results. However, given that the shape and magnitude of the impulse response functions is almost identical to those presented herein we felt that it was appropriate for the sake of brevity to concentrate on the results from the main sample.

3.3 Conclusions

We have outlined and documented a new spectral least squares estimator earlier in chapter 2 for fitting vector autoregressions to ultrahigh-frequency data (the contribution of chapter 2). The estimator makes use of the higher-order auto and cross auto-covariance, co-skewness and co-kurtosis of the order book and models the simultaneous evolution of the mid-price return, update speed, bid ask spread by level, and most importantly the order flow imbalance (measured by the relative oil future volume of contracts asked to contracts bid). We know that through layering a high-frequency trader can shock the order flow imbalance to generate shocks in the mid-price, update speeds and spreads. Impulse response analysis disentangles these effects and we show that for crude oil futures the continuous limit (i.e. the update speed needed before the market ceases to be random) is approximately lower than 10 milliseconds. Hence, a trading algorithm that can adjust quotes at a frequency lower than 10 milliseconds can essentially generate arbitrary profits from this trading mechanism.

As our estimator utilizes only auto and cross-auto-covariance, skewness and kurtosis functions, the data handling requirement is much lighter than attempting to directly fit a VAR model with model implied serial correlation structure to data via maximum likelihood or GMM (the contribution of chapter3) . The spectral estimators can be downshifted in frequency depending on computational capacity. we have evaluated the performance of the spectral least squares approach relative to simple OLS in a Monte-Carlo setting and demonstrated that when the data generating process is contaminated with a complex auto and cross auto covariance, skewness and kurtosis structure the estimator out-performs simple OLS

by a wide margin. In practice this allows the identification of a deterministic lag structure that OLS fails completely to identify.

Chapter 4

Bootstrap eigenvalue correction to test for the number of Latent Factors

4.1 Introduction

Estimating the ex-post quadratic variation of asset prices from high-frequency data is an important tool in asset pricing. Recent results by Dovonon, Goncalves, and Meddahi [2013] have derived the asymptotic and sample characteristics for i.i.d. bootstrapping of a bi-variate covariance matrix from high-frequency returns and compute critical statistics for realized regressions. In this chapter we will demonstrate a bootstrap procedure for covariance matrices of dimension greater than 2 (dimension), and implement a test for rank-deficiency. We then apply this technique to the problem of imputing multivariate hedging ratios from the

cross-section of futures prices measured at ultra-high (millisecond) time frequencies. We demonstrate the robustness of our approach to mild miss-specification and perform an example in sample analysis of the hedging efficacy against the naive long run hedge and a multi-variate GARCH hedge estimated at the daily frequency.

Measuring the ex-post quadratic variation of asset prices at high frequencies (down to the millisecond or even microsecond) has been the subject of considerable recent research. A popular feature of this literature has been to derive critical statistics of interest, such as capital-asset-pricing,beta models from realized regressions.

Indeed, the majority of classical asset pricing models rely on determining the positive definiteness of the co-volatility matrix determining the degree of quadratic variation. However, in most practical applications such as large-scale portfolio management or futures hedging the $k > 2$-dimensional covariance matrix of observed prices is rank-deficient. Subsequent to this, an important task of the econometrician is to determine the degree of rank deficiency and determine the $h < k$ factors driving the quadratic variation of observed prices.

For many types of asset pricing problem, futures markets being an obvious example, the daily term structure of futures prices is presumed to be determined by a very small number of latent factors that determine the difference between the spot and long future and the shape of term structure curve. When speculating and hedging futures positions an important task is to compute the univariate and multi-variate hedging ratios for one or more positions that minimizes the total variance of the portfolio of spot and futures contracts.

The speed of updates in futures trading is often conducted at very high speed, new quotes on limit order books often appear in update times of around one millisecond. Whilst some concern has been voiced that high- frequency and/or algorithmic trading may erode the informational processing capacity of such markets; studies such as Bollen and Whaley [2015] have indicated that whilst the average level of realized volatility of several futures markets has not changed markedly over time, the degree of variation in realized volatility is considerable. As such the use of very high frequency data to build effective hedging ratios is an area of active academic investigation.

Very high frequency data and many of the associated microstructure effects observed at such high frequencies can prove challenging for the computation of variance-covariance matrices, something that has been extensively documented in the equity markets, see Jacod, Li, Mykland, Podolskij, and Vetter [2009b] and Barndorff-Nielsen, Hansen, Lunde, and Shephard [2009b] for extensive univariate examples and, Barndorff-Nielsen and Shephard [2004a] for the simple multivariate case and Barndorff-Nielsen, Hansen, Lunde, and Shephard [2009b] for the multivariate case with asynchronous updating.

Most hedging ratios, both for multi horizon and single horizon hedging, are constructed via inversion of the integrated covariance matrix of futures prices. Consistent estimation of the ex-post time series variation in this matrix provides important information to the hedger/speculator seeking to design a specific position in such a market. Moreover, most treasury management problems involve far more complicated problems than simple pair-wise hedging and a fuller evaluation of the statistical processes driving the complete term structure of futures

is needed.

The rank deficiency of the covariance matrix of multiple futures prices over a given time horizon is implicit. There are usually only a small number of forward looking factors that affect the futures curve. Indeed many models utilize simple two and three factor models that anchor the spot factor a long maturity forward price and a shape factor for the curve. Identification of these factors often uses some form of statistical factor analysis such as principal components or similar. However, the shape of the curve and the number of factors priced in the curve by traders is subjective to the current collective information processing of the market as it clears. There is no specific constraint on how many factors must be included to appropriately model the futures curve.

Modelling the factors requires several directions for filtration. First, there is the time to maturity variation volatility. Samuelson [1965] derived an equilibrium pricing of futures contracts that predicts a monotonic increase in volatility with decreasing time to maturity. However, the instantiation of this effect both in the more recent theoretical literature and from empirical observations of futures prices has been mixed. Hong [2000] provides an overview of the main arguments for and against the maturity effect in futures using a Kyle [1985b] approach with a one-dimensional underlying valuation factor. This approach indicates that whilst hedgers versus speculators (in the classic Keynes framing) are important drivers of maturity effects, informational asymmetries and the mechanics of trading are also important.

Once we have filtered for the volatility component in the cross-sectional variation in futures prices we also have to address the correlation component that

is an emergent property of the underlying pricing factors. Correlation in futures prices in an ideal market is driven by the loadings of the individual, uncorrelated, underlying factors. However, a traded future is subject to both systematic and idiosyncratic microstructure effects. These effects can contaminate the ex-post estimation of realized correlation and potentially bias the identification of the underlying factors.

Our approach and contribution are two-fold. First, we address the statistical identification of the rank of a realized covariance matrix in the presence of microstructure noise and other contaminants via the statistical properties of both the integrated eigenvalues of the spot covariance matrix and the eigenvalues of the integrated covariance matrix via bootstrap. Second, we design and implement a strategy for extracting both the integrated spot eigenvectors and the eigenvectors of the integrated covariance matrix to extract the factors and hence design a mechanism for implementing futures strategies using the ex-post variation in these measures.

4.1.1 Principal component analysis (PCA)

Everyday across the world, thousands of analysts conduct effectively the same exercise: collect their asset data (usually at daily frequency or lower) and compute either a rolling or long run covariance matrix; compute and sort the eigenvalues of this matrix, take the top three (three is a very popular number); extract the corresponding eigenvectors and for the inner product with the original data to form uncorrelated factors. For many applications like yield curve analysis the choice of factors is relatively static, but for forward curves that have 'unusual'

shapes, the given number of forward factors may be unstable. Alternatively, the number of factors is stable, but the weightings and loadings (the eigenvectors and the parameters that relate the factors to the asset prices of interest respectively) maybe time in-homogeneous.

4.1.2 Futures

One application for a PCA type analysis is in the area of futures contracts and in particular assessing the day-to-day number of factors driving the forward-looking prices of these contracts. Futures markets are very actively traded and generate a vast quantity of prices. However, microstructure effects substantially reduce the usefulness of this data for this approach. Hence the majority of futures studies focus on low-frequency data and compute covariance matrices using multivariate GARCH models or other similar tools. However, relatively recent developments in the realized variance-covariance area from high- frequency data can be exploited to provide new insight.

4.1.3 Portfolio management

A second application is using PCA to detect the number of factors in a large cross section of stock returns, such as the S&P 500. For this thesis we will follow Aït-Sahalia and Xiu [2018] and take the 100 most actively traded stocks from the S&P 500 at a five-minute frequency for a selection of days. This is not our core area, but does allow direct comparison with previous work that mainly focuses on equities.

4.1.4 The objective of this chapter

Our primary objective is to have a diagnostic test that establishes the rank of the realized covariance matrix at high frequency in the presence of microstructure noise. To establish how stable the rank, the non-zero latent roots and the weightings are over time,and combine the above into a simulated hedging effectiveness model. As such the model that we see to analyze is as follows:

$$y(t+h) = y(t) + \int_t^{t+h} L(s)f(s)ds + \int_t^{t+h} v(s)z(s)ds \tag{4.1}$$

$$\equiv y(t)ds + \int_t^{t+h} \bar{y}(s) + \int_t^{t+h} \epsilon(s)ds \tag{4.2}$$

where $y(t)$ is a $K \times 1$ vector of log prices, $f(t)$ is a $H \times 1$ vector of driving factors, which are usually assumed to be finite activity cádlág processes, $L(t)$ is a potentially stochastic $K \times H$ matrix of factor loadings and, $v(t)$ is a scalar stochastic volatility parameter determining the degree of microstructure noise across the price process and $z(t+h) - z(t) \sim \mathcal{N}(0, hI)$ determines the microstructure noise. Given this presumed data-generating process, the integrated quadratic variation of $\widetilde{RV} = \int_t^{t+h} y(s)y(s)'ds$, will be of the following form:

$$\widetilde{RV} = \Lambda\Lambda' + \sigma I \tag{4.3}$$

when $< L_{ij}(t), z_k(t) >=< v(t), z_k(t) >=< f_j(t), z_k(t) >=< Lij(t), v(t) >= 0$ and

$$\Lambda = \int_t^{t+h} L(s)ds, \quad \sigma = \int_t^{t+h} v(s)ds \tag{4.4}$$

4.1.5 An informal discussion of the IID boostrap case

To review our approach to the testing problem we will diverge from our focus on oil futures to look at the empirical analysis in Aït-Sahalia and Xiu [2018], which looks at the 100 most actively traded stocks from the S&P 500 cross-section. As this is an illustration we will take one month of data for the S&P 500 on a five-minute grid, excluding the time periods for exchange closures and the first five minutes of trading each day.

The simplest starting point for our bootstrap is the referred to as the IID case, that is, the factors and the noise are driven by IID normally distributed random variables. Hence, $f(t+h) - f(t) \sim \mathcal{N}(0, hI)$ and $\Lambda(t) = \Lambda$, $\forall t$. If we assume that h is a single time increment then for a block of data sampled at times $t_1, t_2, \ldots, t_T$, the data generating process can be written in matrix form as follows:

$$\mathbf{Y} = \mathbf{F}\mathbf{\Lambda} + \sigma\mathbf{Z} = \bar{\mathbf{Y}} + \mathbf{E} \tag{4.5}$$

Each individual column of $\mathbf{Y}$ is formed of T rows of IID normal increments. Hence, the objective is to understand the asymptotic properties of $\mathbf{A} = \mathbf{Y}'\mathbf{Y}$. Notice, that this problem is identical to the standard principal component problem, except that there is noise within $\mathbf{E}$. As such, the rank of $\text{rk}(\mathbf{Y}'\mathbf{Y}) = K$. Indeed, the objective of the analysis is not to identify the rank of $\mathbf{Y}'\mathbf{Y}$ but of $\bar{\mathbf{Y}}'\bar{\mathbf{Y}}$. However $\bar{\mathbf{Y}}'\bar{\mathbf{Y}}$, is unobserved. So we will demonstrate asymptotically how the standard technique for PCA will yield the uncorrelated factor structure then outline how an IID bootstrap consistently estimates the factor structure. we then move on to a more conventional setting with a Levy type process, but in this case

with a fixed time horizon, hence the complication of the following limit:

$$\lim_{T\to\infty} T^{-1}\mathbf{Y}'\mathbf{Y} = \mathbf{0}, \quad \text{Realized Variance Limit} \tag{4.6}$$

where $\mathbf{0}$ is a $K \times K$ null matrix, which is in contrast to

$$\lim_{T\to\infty} T^{-1}\mathbf{Y}'\mathbf{Y} = \mathbf{Q}, \quad \text{Infinite Time Horizon Limit} \tag{4.7}$$

for the classical limit in econometrics.

4.1.6 Asymptotic properties of $\mathbf{Y}'\mathbf{Y}$ under the classical limit theorem

It only requires a small adjustment to the standard tests Anderson [1959], Bartlett [1963], and Muirhead [1982] to develop the base case asymptotic theory for $\mathbf{Y}'\mathbf{Y}$ and hence develop a likelihood ratio test for the rank of $\bar{\mathbf{Y}}'\bar{\mathbf{Y}}$. Our analysis will proceed using the Hermite polynomial approach of [Muirhead, 1982].

However, the analysis of Anderson [1959] using partitioning provides an identical result.

Our analysis will focus on the following likelihood ratio statistic:

$$\mathfrak{L}_k = -n \log V_k$$

where:

$$V_k = \frac{\prod_{i=k+1}^{p} l_i}{1/(p-k)(\sum_{i=k+1}^{p} l_i)^{p-k}}$$

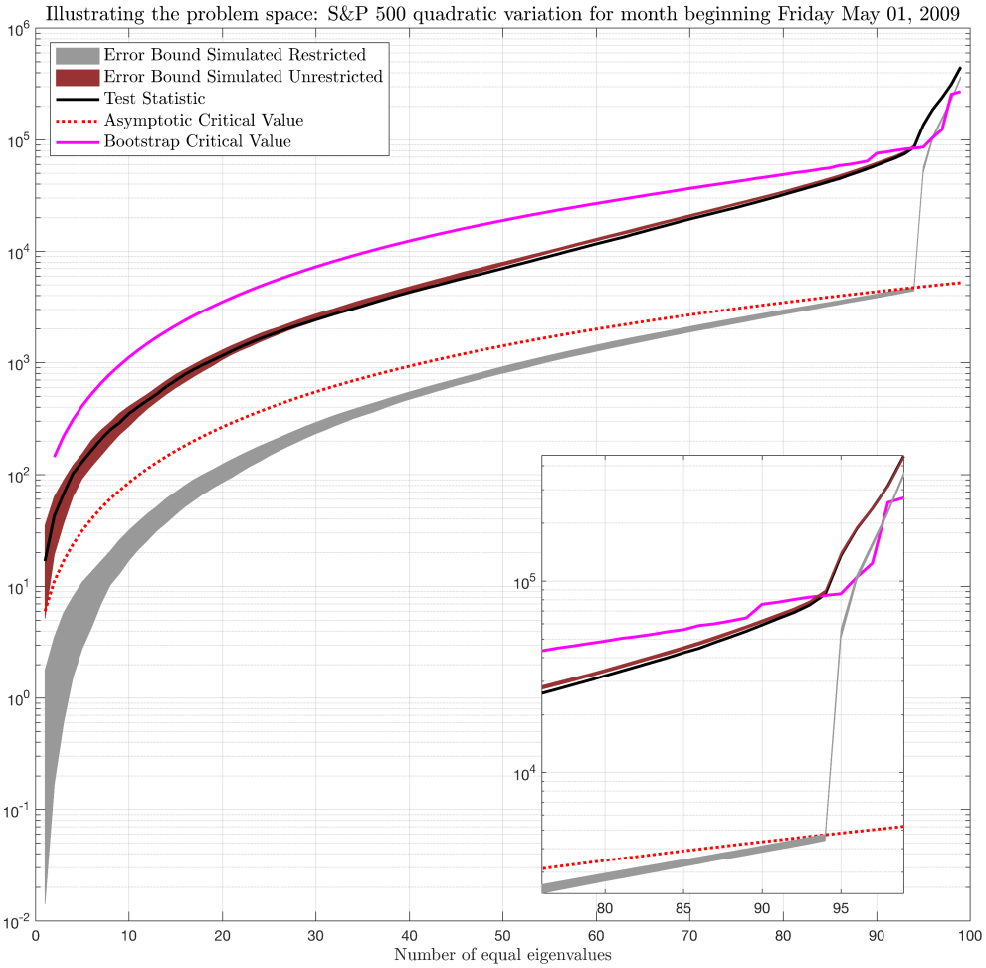


Figure 4.1: Illustration of the correction problem.
In this figure we examine the testing problem for S&P 500 data for one month for in five minutes grid, to show how the noise effect the data under the test statistics.

All of our tests in this chapter we will use this statistic. Here p is the dimension and $q = p - k$ is the number of bounded eigenvalues.

Figure 4.1 illustrates the testing problem. We collected data for one month from the S&P 500 cross-section. We have sorted the stocks by most actively traded (that is the number of updated price change to the mid-price) and then made a five-minute grid. The unbroken black line is the test statistic, the abscissa (x-axis) values represent the number of equal eigenvalues. Hence the far right hand side would indicate that the data are formed of 100 identically distributed noise factor and the far left 100 uncorrelated but with differente variance factors. The statistic is the same for all models and computed from the data, for a given critical statistic the point at which the critical value crosses the test statistic (going from right to left, this is from below to above), is the test statistics indicated number of factors. The objective is to gain insight into the factor structure of the data using this test statistic. Eyeball inspection clearly shows that there is a kink when we presume that $\lambda_7 = \lambda_8 \ldots = \lambda_1 00$, that is we might infer from visual inspection that the first 94 eigenvalues are bounded and equal, indicating six (100 - 94) factors explaining the data. However, the critical value for the traditional test statistic, shown as dashed red line, is substantially below the test statistic for all hypothesized bound eigenvalues. This suggests that the factor covariance matrix is full rank and NONE of the eigenvalues are bounded and identical. Is it possible to see if the test statistic would fail to reject for more than six heterogeneous eigenvalues? The answer is yes. If we take the estimated

covariance matrix $\boldsymbol{S}$ for this day and decomposes into:

$$\boldsymbol{S} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}'$$

where $D = \text{diag}[\boldsymbol{l}]$ is the diagonal vector of eigenvalues and $\boldsymbol{V}$ is a matrix with columns formed from the eigenvectors of $\boldsymbol{S}$ corresponding to each estimated eigenvalue $l_i \in \boldsymbol{l}$, sorted from largest to smallest (we sort the order of the stocks in this way for ease of comparison). If we impose the supposed eigenvalue structure:

$$l_7 = l_8 = \ldots = l_{100} = \text{average}[l_7 \ldots l_{100}]$$

by setting $l^*_{i\in 7,\ldots,100} = 1/94 \sum_{i=7}^{100} l_i$, whilst keeping $l^*_i = l_i$ for $i \in 1, \ldots, 6$ we can create a new covariance matrix $\boldsymbol{S}^* = \boldsymbol{V}\boldsymbol{D}^*\boldsymbol{V}'$ where $\boldsymbol{D}^* = \text{diag}[\boldsymbol{l}^*]$. Generating a new dataset $X^* \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}^*)$, we can recompute the test statistic under the null that the 94 smallest eigenvalues are bounded and recompute the test statistic for $k = 0$ to $k = 100$ (k is the factors). Repeating this calculation 1000 times we generate the area shaded in grey in 4.1. We can see that this set of test statistics cross the red line at 94, as anticipated as the normally distributed and generated under the null. Hence, if the data is normal the classic statistic will return the factor structure we anticipate. By comparison we simulate the data under the alternative to compute the dark red area to illustrate the opposite effect. However, we need to correct for the inherent bias in estimating l_i from data, that is, the distribution of the eigenvalues recovered from the stock price data will not match that anticipated from a normally distributed random variable. Hence, we need to estimate the empirical distribution of the eigenvalues from the bias inherent in

estimating them. For this we choose a bootstrap device and dynamically recompute the null hypothesis. This is illustrated in what follows. However, you can see the effect of simulating the empirical distribution of the eigenvalues, by looking at the critical statistic computed in unbroken purple Figure 4.1, this is significantly higher (to match the high variability in the bounding of the eigenvalues in the sample covariance matrix) and crosses the black line (from below to above) at 94 =100 - 6, which is a similar point to our eyeball estimate from the pattern of the statistic. Hence, we can infer that six factors is indeed a reasonable presumption. In follows we will go through in detail, how the classical bound (the red dotted line) is computed and then demonstrate how we adjust the bound to recompute the new critical statistic (illustrated in purple).

The intuition of this experiment underpins how the bootstrap estimator works. First, generate the classic likelihood ratio test first proposed in Anderson [1963] then adjust the critical bounds to account for the fact that the presumed data generation is non-gaussian. To illustrate the consistency of the bootstrap, we will show that the likelihood ratio test under the Gaussian assumption can be consistently recovered through the simulation exercise above and then move on to more complex simulation conditions with jumps and stochastic volatility.

4.2 Proof of consistency

The proof sketch of consistency for our preferred bootstrap device outlined later on, is as follows:

- Write down the classical MLE estimator for the multivariate Gaussian case.This is from Anderson [1963], but we will use the approach to the derivation in [Muirhead, 1982].

- Next determine the likelihood ratio under the null and alternative for a stepwise test, as the number of proposed factors goes from $k = 0$ to $k = p-1$.

- Write down the distribution under the null and alternatives. Anderson [1963] uses a Hilbert space for the set of applicable matrices under these conditions and then block partitions it. Muirhead [1982] directly invokes spectral matrix theory to compute the sums of the estimated eigenvalues and this is closer to the approach in [Aït-Sahalia and Xiu, 2018].

  Hence we are going with this approach.

- Once we write down the likelihood ratio case in a Neyman-Pearson style case as a chi squared distribution, for each step from $k = 0$ to $k = p - 1$ factors, we will show that simulating directly the restricted matrix with bias adjusted eigenvalues returns the same test statistic distribution as the asymptotic theory under the null and alternatives.

- Finally, we adjust the data generating process to a more realistic case and then recompute the test statistic under this case. As the asymptotic distribution of the statistic is impossible to identify under Aït-Sahalia and Xiu

[2018], without knowing the factor structure prior to the test we will use this simulation to reconstruct the test statistic distribution under the null and then compare to the observed statistic. we conduct size and power tests on sample data to confirm our results. Future work can focus on particular simple distributional cases to see if an asymptotic theory can be built to confirm the prima facie evidence from our simulation analysis.

4.2.1 Notation and the classical MLE estimator of the sample covariance

We will set out the problem of determining the statistical properties of the roots of $\mathbf{A} = \mathbf{Y}'\mathbf{Y}$ by bootstrap, under its functional analogue $A = Y'Y$, where Y is an $n \times m$ matrix (equivalent to $T \times K$ for our general data sampled from a Levy process in $\bar{\mathbf{Y}}$). Let $S = n^{-1}A$ be the sample covariance matrix. We know that if $Y \sim \mathcal{N}(0, \Sigma \otimes I_n)$, then A is Wishart distributed, with n, m degrees of freedom, that is:

$$A \sim \mathcal{W}(n, m, \Sigma) \tag{4.8}$$

where:

$$\mathfrak{F}(A) = \frac{1}{2^{\frac{mm}{2}} \Gamma_m(\frac{1}{2}n)(\det \Sigma)^{\frac{1}{2}}} \exp\left(\mathrm{tr}(-\tfrac{1}{2}\Sigma^{-1}A)(\det A)^{\frac{n-m-1}{2}}\right) \tag{4.9}$$

where $\Gamma_m(a)$ is the multivariate Gamma function,

$$\Gamma_m(a) = \pi^{\frac{m(m-1)}{4}} \prod_{i=1}^{m} \Gamma(a - \tfrac{1}{2}(i-1)) \tag{4.10}$$

and $\Gamma(c) = (c-1)!$ is the univariate gamma function and ! is the factorial operator. Also, F is the function. Our proof of consistency for the bootstrap is outlined in the following stages:

1. Write down the exact maximum likelihood estimator for $S = (n-1)^{-1}A$.
2. Determine the exact cumulant of A under the multivariate normality assumption $Y \sim \mathcal{N}(0, \Sigma \otimes I_n)$.
3. Determine the probability distribution for the latent roots of S.
4. Write down the likelihood ratio test, in terms of Wishart densities, of Anderson [1963] in cumulant form.
5. Use an Edgeworth expansion to approximate the cumulant.
6. Illustrate that the IID bootstrap cumulant is identical to the Edgeworth expansion of the likelihood ratio and hence approximates the quantity.

Each stage is set out as a proposition (6 propositions), with a corresponding proof. The first three are well known, hence we do not provide a detailed proof, just a summary. For more details see Muirhead [1982] amongst others. Steps four and five are new and we will review them in some detail.

Proposition 1. *The Sample Maximum Likelihood Estimator for* Σ

The starting assumption is that $Y \sim \mathcal{N}(0, \Sigma \otimes I_n)$*, where* Σ *is presumed to be positive definite and hence full rank. Anderson [1959] proves that the sample maximum likelihood estimator of the covariance matrix* $cov[Y]$ *is* $n^{-1}Y'Y$*. We can illustrate this as follows:*

Proof. Proof of Proposition 1 If Y is not centred then for a row of Y denoted Y_i, note that the standard estimator is

$$A = \sum_{i=1}^{N} (Y_i - \bar{Y})(Y_i - \bar{Y})' \tag{4.11}$$

where $N = n - 1$ and $\bar{Y} = N^{-1} \sum_{i=1} Y_i$ is the vector of means for each element of Y_i. Let $\boldsymbol{\mu}$ be the true mean, hence:

$$A = \sum_{i=1}^{N} ((Y_i - \boldsymbol{\mu}) - (\bar{Y} - \boldsymbol{\mu}))((Y_i - \boldsymbol{\mu}) - (\bar{Y} - \boldsymbol{\mu}))' \tag{4.12}$$

Rearrange to remove the invariant elements from the sum:

$$A = \sum_{i=1}^{N} (Y_i - \boldsymbol{\mu})(Y_i - \boldsymbol{\mu})' - N(\bar{Y} - \boldsymbol{\mu})(\bar{Y} - \boldsymbol{\mu})' \tag{4.13}$$

Taking expectations:

$$\mathbb{E}[A] = \sum_{i=1}^{N} \mathbb{E}[(Y_i - \boldsymbol{\mu})(Y_i - \boldsymbol{\mu})'] - N\mathbb{E}[(\bar{Y} - \boldsymbol{\mu})(\bar{Y} - \boldsymbol{\mu})'] \tag{4.14}$$

From the properties of the normal distribution that $\mathbb{E}[(Y_i - \boldsymbol{\mu})(Y_i - \boldsymbol{\mu})'] = \Sigma$ and $\mathbb{E}[(\bar{Y} - \boldsymbol{\mu})(\bar{Y} - \boldsymbol{\mu})'] = \frac{1}{N}\Sigma$, therefore:

$$\mathbb{E}[A] = \sum_{i=1}^{N} \mathbb{E}[(Y_i - \boldsymbol{\mu})(Y_i - \boldsymbol{\mu})'] - N\mathbb{E}[(\bar{Y} - \boldsymbol{\mu})(\bar{Y} - \boldsymbol{\mu})'] \tag{4.15}$$

$$\mathbb{E}[A] = N\Sigma - N\frac{1}{N}\Sigma \tag{4.16}$$

$$\mathbb{E}[A] = (N - 1)\Sigma \equiv n\Sigma \tag{4.17}$$

Now it remains to show that for a Wishart distribution $\mathcal{W}(n, m, \Sigma)$, the sample log likelihood for A is maximized at:

$$A = \sum_{i=1}^{N} (Y_i - \boldsymbol{\mu})(Y_i - \boldsymbol{\mu})' - N(\bar{Y} - \boldsymbol{\mu})(\bar{Y} - \boldsymbol{\mu})' \tag{4.18}$$

such that for $nS = A$, S is a consistent estimator of Σ. This is given in Theorem 3.2.3 in page 91 of Muirhead [1982] and excluded for reasons of compactness. ■

We design a bootstrap algorithm specifically to test for numbers of latent roots (and hence the rank and optimal number of components) specifically designed for very high frequency data. We currently have one version of algorithm complete, an IID bootstrap extracting the eigenvalues integrated realized covariance matrix. A second version using a en-block bootstrap following recent asymptotic work by [Aït-Sahalia and Xiu, 2018].

We then utilize a novel order-book dataset to implement the algorithm to compute hedging ratios from daily updates of the implied optimal numbers of factors. We have prima-facie evidence to suggest that this is a more effective mechanism for determining the optimal hedging ration than a standard GARCH on daily data alone or a naive hedge (albeit it is a little bit of an unfair comparison). In principle it could be possible to aggregate multiple time frequency versions of our model (say at 100ms, 1second, 5second, 30second, 1minute, 5minute) to construct smoothing kernels by factor to unpick the microstructure noise from the equilibrium prices, we leave this to future work. The focus of this chapter is methodological and motivation, so you can review our empirical results and consistency proofs later in this chapter.

4.3 Related Work

4.3.1 Earlier work on futures, PCA and RV

The obvious immediately comparable research to this one is Aït-Sahalia and Xiu [2018] . The authors use spectral functions to derive an estimate of the asymptotic distribution of the latent roots of the covariance matrix. They then derive two approaches, one to compute the eigenvalue of the integrated covariance matrix (directly analogous to our approach) and their second, preferred approach, to compute the integrated eigenvalues from the spot covariance matrix. The extracted eigenvectors are then used to compute high-frequency daily factors for a cross section of 100 US stocks. In terms of our bootstrap methodology the nearest comparator is Dovonon, Goncalves, and Meddahi [2013] , who derive the asymptotic consistency of an IID bootstrap for bivariate and high realized variance-covariance matrices. We will illustrate the analogous bootstrap methodologies for the two proposed asymptotic approaches of Aït-Sahalia and Xiu [2018]. However, our current implementation is limited to analysis of the integrated covariance matrix in the spirit of Dovonon, Goncalves, and Meddahi [2013], Jacod, Li, Mykland, Podolskij, and Vetter [2009a] and [Barndorff-Nielsen, Hansen, Lunde, and Shephard, 2009b].

4.3.2 The data-generating process

Let t, for $t \geq 0$ denote the K vector of equilibrium log spot and futures prices over a single day of trading, where $\tilde{p}_k(t)$ is the equilibrium price of the k element and $\mathcal{T}_k$ is the dated maturity for each element k. Let $\tau_k = \mathcal{T}_k - t$ represent the tenor of the k price in this vector, such that $\tau_1 \to 0$, i.e. the spot rate is the effective immediate delivery and $\tau_{k+1} > \tau_k$ as such the futures prices are ordered by tenor from shortest to longest. we will assume that $0 = \mathbf{0}_K$ such that t is in effect a cumulative return from 0 to t, we assume one trading day is over the interval $[0, 1]$. As our interest is just one day, we ignore the drift term and focus exclusively on volatility, hence: $\mathrm{d}t = \mathrm{d}\tilde{\Lambda}(t)\mathrm{d}\tilde{W}(t)$.

4.3.3 Reducing the rank

Where $\tilde{\Lambda}(t)\tilde{\Lambda}(t)' = \tilde{\Omega}(t)$ denotes the matrix of instantaneous covariation or cross products between assets and $\tilde{W}(t + \Delta t) - \tilde{W}(t) \sim \mathcal{N}(\mathbf{0}_H, \Delta t \mathbf{I}_H)$ are independent Weiner processes of dimension $H \leq K$ and $\mathbf{I}_H$ is a H dimensional identity matrix. $\tilde{\Lambda}(t)$ is a $K \times H$ matrix and the rank $K - H$ physical spot covariance matrix is $\tilde{\Omega}(t)$ is a $K \times K$ matrix. However, we do not observe the equilibrium log spot and futures prices, we follow a price $p(t)$ that is a function of the noisy transaction environment. Accordingly we presume that another full-rank noise component is present in the data. Hence:

$$t = t + noise(t),$$

we impose some structure on the noise later on, but our main assumptions is that it's quadratic variation is of full rank.

4.4 High-Frequency PCA: ideas and data

4.4.1 Notations

At any given time t each level, indexed by j, reports three pieces of information: The price $P_j^{\mathscr{Q}}(t)$, where $\mathscr{Q} \in \{\mathscr{B}, \mathscr{A}\}$. The volume of contracts associated with each quote $V_j^{\mathscr{Q}}(t)$. The number of active trading accounts associated with those quotes $N_j^{\mathscr{Q}}(t)$. Note that the ratio $V_j^{\mathscr{Q}}(t)/N_j^{\mathscr{Q}}(t)$, gives a measure of concentration of quotes per active buyer or seller. Whilst the order-book is a standing entity, it is useful to characterize a 'refresh-time' for the order-flow. Let $\mathbf{t}$ represent the M length vector master update sequence for all futures contracts and $\mathbf{t}_k$ be the individual clocks for the k contract. Later it will be useful to specify $\mathbf{t}_j^{\mathscr{Q}}$ and $\mathbf{t}_{jk}^{\mathscr{Q}}$, to represent the update times for individual components indexed by $j \in \{1, \ldots, J\}$ level for the $k \in \{2, \ldots, K\}$ futures contract or spot price (when $k = 1$). We can then index a day by $m \in \{1, \ldots, M\}$ updates, Each update has a potentially idiosyncratic time stamp t_m.

4.4.2 Important point

As our objective is to determine the number of principal components, we are almost uniquely vulnerable to the Epps effect. Standard likelihood ratio based tests are computed from correlation matrices and underestimation of the magnitude of correlations will markedly reduce the discriminatory of our test statistics. Hence we approach the problem from two directions. First, we use bootstrap to improve the discriminatory power of our test, and second, we make the most of the market data available to ensure that stale prices and updating is less problematic.

4.4.3 Things to consider

As we move to much higher frequencies, circa 100s of milliseconds and higher, we need to keep in mind the following points. The longest (and to an extent the shortest) maturity data we can consider will probably be constrained. Indicatively, from looking at Crude Oil, Eurodollars, S&P 100 and S&P 500 E-mini contracts we can cope with, from around 5 years to about a month, with the spot generally taken as the shortest tenor future. Moving to the very highest frequency of updates – timescales around 1 millisecond to 50 milliseconds, we are also constrained by time of day. Indeed, theory suggests that time of day needs to be controlled for Admati and Pfleiderer [1988] and [Admati and Pfleiderer, 1989].

Moreover, to high time-scales involve far greater levels of microstructure noise compared to lower frequencies, but the overall information set, should hopefully be higher. Aït-Sahalia and Xiu [2018] give a very useful summary on this trade-off.

4.4.4 What is microstructure noise in this context?

At tick level, return autocorrelations are often highly negative. Additionally, there is often highly significant, covariation between variation in the relative volume of the bid and ask side of the order book and mid-price changes. Given that each tenor has its own order-book, this in part motivates our assumption of the microstructure noise being full rank and possibly independent.

4.5 Overview on the tests

4.5.1 Constructing the time-matched data matrix

Let $\mathbf{P} = [\mathbf{p}_k^*]_{k\in\{1,\dots,K\}}$, where $\mathbf{p}_k^*$ is the transformation of $\mathbf{p}_k$ to a single master grid given by our master clock $\mathbf{t}$. Choosing $\mathbf{t}$: there are several obvious approaches: A uniform time clock from a specified start time $t_{m=1} = 0$ to a specified end point $t_{m=1} = 1$, gives $1/M$ fractions of a trading day. Choose a master clock from the set of clocks $\mathbf{t}_{k\in\{1,\dots,K\}}$, for instance the least frequently updated contract. Follow Barndorff-Nielsen, Hansen, Lunde, and Shephard [2009b] Audrino and Corsi [2008] and Aït-Sahalia and Xiu [2018] ensure that the estimated covariance matrix is positive definite by forcing the number of columns to be less than the dimension of the asset prices. For a bootstrap purposes will remove the forward 'jittering' (the random resampling of the start times) as this is implicit in the bootstrap device.

4.5.2 The generic design of tests

Let $\mathbf{R} = [\Delta p_{mk}^*]_{k\in\{1,\dots,K\},m\in\{1,\dots,M\}}$ be the $N-1\times K$ matrix of returns. our presumption with be that:

$$\mathbf{R} = \tilde{\mathbf{R}} + \tilde{\mathbf{Z}}$$

Where $\tilde{\mathbf{R}}$ is the rank deficient equilibrium log price change and $\tilde{\mathbf{Z}}$ is the full rank microstructure noise. Aït-Sahalia and Xiu [2018], propose two approaches for deducing the optimal rank and hence the eigenvalues and eigenvectors of interest. Decompose $\mathbf{R}$ into non-overlapping blocks $\mathbf{R} = [\mathbf{R}_b']_{b\in\{1,\dots,B\}}'$. For each block $\mathbf{R}_b$

compute the integrate covariance matrix $\mathbf{C}_b$ using a jump robust approach (they suggest an exclusion of rows with elements above a certain absolute magnitude). Compute the 'spot' eigenvalues and aggregate them. This is referred to as the 'integrated eigenvalues of the spot covariance matrix'.

Compute the integrated covariance matrix $\mathbf{C}$ directly from $\mathbf{R}$ and compute the eigenvalues of this matrix. This is referred to as the 'eigenvalues of the integrated covariance matrix'. Aït-Sahalia and Xiu [2018] make a good case that the integrated eigenvalues derived from aggregating the eigenvalues of $\mathbf{C}_b$ are more useful when the spot covariance process is instantaneously stochastic. Aït-Sahalia and Xiu [2018] impute asymptotic the distributions of the terminal eigenvalues using a spectral approach and compared the estimates to the asymptotic distributions to infer the optimal number of principal components.

4.5.3 The design of tests

our approach to testing is a little bit simpler, but we will make use of a bootstrap device to attempt to correct for the power and size problems induced by the potential variation in $\tilde{\mathbf{Z}}$, we propose as follows: First, compute $\mathbf{S}^{\dagger\dagger} = \mathbf{R}'\mathbf{R}$ and normalize by the diagonal elements,

$$\mathbf{S}^{\dagger} = diag[diag[\mathbf{S}^{\dagger\dagger}]^{-\frac{1}{2}}]\mathbf{S}^{\dagger\dagger}diag[diag[\mathbf{S}^{\dagger\dagger}]^{-\frac{1}{2}}]' \tag{4.19}$$

to recover the integrated correlation matrix. Then compute the estimated eigenvalues of $\mathbf{S}^{\dagger}$, denoted $\ell = \{\ell_1, \ldots, \ell_K\}$ and sort them from largest to smallest, ensuring that the collection of eigenvectors $\mathbf{l}_{k\in\{\ell_1,\ldots,\ell_K\}}$ preserves this sorting. Re-

call that for a correlation matrix the sum of the eigenvalues on the diagonal is equal to the dimension of the matrix. Sequentially compute the following statistic K times:

$$\begin{aligned}\mathscr{V}_k &= -N \log[V_k] \\ V_k &= \frac{\prod_{i=k+1}^{K} \ell_i}{\left(\frac{1}{K-k} \sum_{i=k+1}^{K} \ell_i\right)^{K-k}}\end{aligned}$$

This follows the approach of Jacod, Li, Mykland, Podolskij, and Vetter [2009b] and is based on the foundational work of Bartlett in a series of mongraphs from the 1950s.

4.5.4 Classical critical values for $\mathscr{L}_k$

The classical critical values (normal distribution) assume the following distributional properties for $\mathbf{R}$:

$$\mathbf{R} \sim \mathcal{N}[\mathbf{0}_{M\times K}, \Omega \otimes \mathbf{I}_M], \quad \Sigma = \Lambda\Lambda'$$

where Λ is a $K \times H$ real matrix. The noise is asumed to be full rank of the form either:

$$\begin{aligned}\tilde{\mathbf{Z}} &\sim \mathcal{N}[\mathbf{0}_{M\times K}, \Sigma \otimes \mathbf{I}_M], \text{ or} \\ \tilde{\mathbf{Z}} &\sim \mathcal{N}[\mathbf{0}_{M\times K}, \sigma^2 \mathbf{I}_{K\times M}],\end{aligned}$$

where $\Lambda\Lambda'$ and Σ or σ are normalized such that $\mathbf{R}'\mathbf{R}$ is symmetric with a unit diagonal. $diag(\Lambda\Lambda' + \Sigma) = \mathbf{1}_K$ of course proceeds under the assumption that the microstructure noise and the signal are uncorrelated, which we acknowledge is a slightly harder assumption to support. Under the null hypothesis that the 'true' integrated eigenvalues $\lambda = \{\lambda_1, \ldots, \lambda_K\}$ are from a full rank covariance (where λ_K the number of eigenvalues) matrix, the diagonal elements will comprise of K unit variance, therefore $H_0 : \lambda_1 = \cdots = \lambda_K$. The alternative is that $H_k : \lambda_{k+1} = \cdots = \lambda_m, \quad (= \lambda, unknown)$. Setting $q = K - k$, then $-N \log[V_k] \sim \chi^2_{(q+2)(q-1)/2}$, under our normality assumption. It is worth noting that the issue of low power for this statistic is well known see [Anderson, 1963].

Indeed, Bartlett proposed a correction for short samples under normality of

$$\mathscr{V}_k = -\left(n - k - \tfrac{2q^2+q+2}{6q}\right) \log[V_k] \sim \chi^2((p+2)(p-1)/2)$$

Where (p) is the price. This correction works very well under-normality, but still lacks power when the signal-to-noise ratio is high.

4.5.5 Some nice interpretations

The privouse distributional assumption is obviously difficult to support (in 4.5.4) so we move away from it. However, there is a nice intuition to the value of σ. For the case that $\tilde{\mathbf{Z}} \sim \mathcal{N}[\mathbf{0}_{M\times K}, \sigma \otimes \mathbf{I}_{K\times M}]$, asuming that $\text{var}[r_{i,k}] = 1$, then σ is the ratio of full rank microstructure noise to reduced rank signal. As σ increases, then by construction the ratio of full rank microstructure noise increases relative to the reduced rank fluctuations in the signal component. In our simulations

we will allow the microstructure noise DGP to simply have $\mathbb{E}[\tilde{z}_{mk}^2] = \sigma^2$. As suggested specification is that $\tilde{z}_{mk}$ is a stationary AR(1) process with $\mathbb{E}[\tilde{z}_{mk}^2]$ and AR parameter ρ.

4.6 Prior results on the Latent Roots of sample covariance matrix

In this section will recall the mathematical work on the eigenvalues of a sample covariance matrix conducted by Anderson [1959, 1963] and used in Muirhead [1982] book to build this chapter tests. The objective here is to provide the minimum mathematical prerequisites to derive the consistency of the estimator.

The joint density function of the latent roots $l_1, \ldots, l_m$ of the sample covariance matrix S given by Theorem 1 in Anderson [1963] involves the hypergeometric function ${}_0F_0^{(m)}\left(-\frac{1}{2}nL, \Sigma^{-1}\right)$ having an expression in terms of zonal polynomials, If n is large, this zonal polynomial series converges very slowly in general. Moreover, it is finical to realize from this series any impression for the behavior of the density function or an understanding of how the sample and population roots interact with each other.

The zonal polynomial expansion of ${}_0F_0^{(m)}$ does not lend itself easily to the derivation of asymptotic results, hence Muirhead [1982] takes an alternative route to the same result in Anderson [1963], that a suitably bias corrected average of the $p-k$ can be used to establish a bound. Integral representations are generally the most useful tool for obtaining asymptotic results in this type of analysis, so

that here we will work with the integral form:

$$ {}_0F_0^{(m)}\left(-\frac{1}{2}nL, \Sigma^{-1}\right) = \int_{O(m)} etr\left(-\frac{1}{2}n\Sigma^{-1}HLH'\right)(dH) \tag{4.20} $$

and examine its asymptotic behaviour as $n \to \infty$. To do this we will make use of the following theorem which gives a multivariate extension of Laplace's method for obtaining the asymptotic behaviour integrals, this is given in the form expressed in Apostol [1969, Chapter 13]. In this theorem, and subsequently, the notation "$a \sim b$ for large n" means that $a/b \to 1$ as $n \to \infty$.

Theorem 1. *[Muirhead, 1982].*

Let D be a subset of $\mathbb{R}^p$ and let f and g be real-valued functions on D such that:

(i) f has an absolute maximum at an interior point ξ of D and $f(\xi) > 0$;

(ii) there exists a $k \geq 0$ such that $g(x)f(x)^k$ is absolutely integrable on D;

(iii) all partial derivatives

$$ \frac{\partial f}{\partial x_i} \text{ and } \frac{\partial^2 f}{\partial x_i \partial x_j} \ (i, j = 1, \ldots, p) \tag{4.21} $$

exist and are continuous in a neighborhood $N(\xi)$ of ξ;

(iv) there exists a constant $\gamma < 1$ such that

$$ \left|\frac{f(x)}{f(\xi)}\right| < \gamma \text{ for all } x \in D - N(\xi) \tag{4.22} $$

(v) g is continuous in a neighborhood of ξ and $g(\xi) \neq 0$. Then, for large n,

$$\int_D [f(x)]^n g(x)dx \sim \left(\frac{2\pi}{n}\right)^{p/2} \frac{[f(\xi)]^n g(\xi)}{[\triangle\ (\xi)]^{1/2}} \tag{4.23}$$

where $\triangle\ (\xi)$ denotes the Hessian of $-\log f$, namely,

$$\triangle\ (\xi) = det\Omega(\xi), \Omega(\xi) = \left(\frac{-\partial^2 \log f(\xi)}{\partial\xi_i \partial\xi_j}\right) \tag{4.24}$$

end of Theorem 1

The simple idea in the proof contains the determinations for the large n the major contribution to the integral will arise from a neighborhood of ξ and expanding f and g about ξ. We sketch a heuristic proof in the same form as [Anderson, 1959].

First we write

$$\int_D [f(x)]^n g(x)dx = [f(\xi)]^n \int_D g(x) \exp\{n[logf(x) - logf(\xi)]\} dx \quad (4.25)$$

In a neighborhood $N(\xi)$ of ξ, $logf(x)$-$logf(\xi)$ is approximately equal to $-\frac{1}{2}(x-\xi)'\Omega(\xi)(x-\xi)$, $g(x)$ is approximately equal to $g(\xi)$ and then, using (iv), n can be chosen sufficiently large so that the integral over $D - N(\xi)$ is negligible and hence the domain of integration can be extended to $\mathbb{R}^p$. Thus for large n,

$$\int_D [f(x)]^n g(x)dx \sim [f(\xi)]^n g(\xi). \int_{\mathbb{R}^p} \exp\left[-\frac{1}{2}n(x-\xi)'\Omega(\xi)(x-\xi)\right]$$

$$= \left(\frac{2\pi}{n}\right)^{p/2} \frac{[f(\xi)]^n g(\xi)}{[\Delta(\xi)]^{1/2}}$$

The end of Theorem 1 proof.

Theorem 2. *[Muirhead, 1982].*

If $\Sigma = diag(\lambda_1, \ldots, \lambda_m)$ *and* $L = diag(l_1, \ldots, l_m)$, *where* $\lambda_1 > \cdots > \lambda_m > 0$ *and* $l_1 > \cdots > l_m > 0$ *then, for large* n,

$$ {}_0F_0^{(m)}\left(-\frac{1}{2}nL, \Sigma^{-1}\right) \sim \frac{\Gamma_m\left(\frac{1}{2}m\right)}{\pi^{m^2/2}} \exp\left(-\frac{1}{2}n\sum_{i=1}^{m}\frac{l_i}{\lambda_i}\right)\prod_{i<j}^{m}\left(\frac{2\pi}{nc_{ij}}\right)^{1/2} \tag{4.26} $$

where

$$ c_{ij} = \frac{(l_i - l_j)(\lambda_i - \lambda_j)}{\lambda_i\lambda_j} \tag{4.27} $$

The approach identified in Anderson (1959) is to write the ${}_0F_0^{(m)}$ *function as a multiple integral, hence we write:*

$$ {}_0F_0^{(m)}\left(-\frac{1}{2}nL, \Sigma^{-1}\right) = \int_{O(m)} etr\left(-\frac{1}{2}n\Sigma^{-1}HLH^{'}\right)(dH) \tag{4.28} $$

Here (dH) *is the 'normalized invariant measure' on the polynomial* $O(m)$*; Muirhead [1982] takes the formulation of Anderson (1959) then extracts the normalized measure to work with 'un-normalized' measures, this is then expressed in 'big Wedge' outer products as follows:*

$$ (H'dH) = \bigwedge_{i<j}^{m} h_j^{'}dh_i \tag{4.29} $$

As noted by Muirhead [1982] this is equivalent to the ordinary Lebesgue measure as a point set in Euclidean space of dimension $\frac{1}{2}m(m-1)$. *These two measures are related by the following formulation*

$$ (dH) = \frac{\Gamma_m\left(\frac{1}{2}m\right)}{2^m\pi^{m^2/2}}I(n) \tag{4.30} $$

so that

$$ {}_0F_0^{(m)}\left(-\frac{1}{2}nL, \Sigma^{-1}\right) = \frac{\Gamma_m\left(\frac{1}{2}m\right)}{2^m \pi^{m^2/2}} I(n) \tag{4.31} $$

where

$$ I(n) = \int_{O(m)} etr\left(-\frac{1}{2}n\Sigma^{-1}HLH^{'}\right)(H^{'}dH) \tag{4.32} $$

Note that this integral has the form

$$ I(n) = \int_{O(m)} [f(H)]^n (H^{'}dH) \tag{4.33} $$

where

$$ f(H) = etr\left(-\frac{1}{2}n\Sigma^{-1}HLH^{'}\right) \tag{4.34} $$

This is the end of Theorem 2*, as presented by [Muirhead, 1982] and provides the base result on what the variation in the eigenvalue structure will look like in an un-normalized statistic. Hence we can now begin to identify some distributional properties of this object under different inherent structures of* Σ*. The innovation in Anderson (1963) is that the bound on the structure is explicit to point-wise entries in* Σ*. In Figure 4.2 we vary the* $\mathcal{L}_2$ *norm of* $f(H)$ *over a single point-wise change in* Σ *and illustrate the single bounded region for which an Neyman-Pearson Likelihood ratio type test is value (around one). Indeed, this illustrates why the test breaks down if the pointwise entry has more anticipated variation than this domain.*

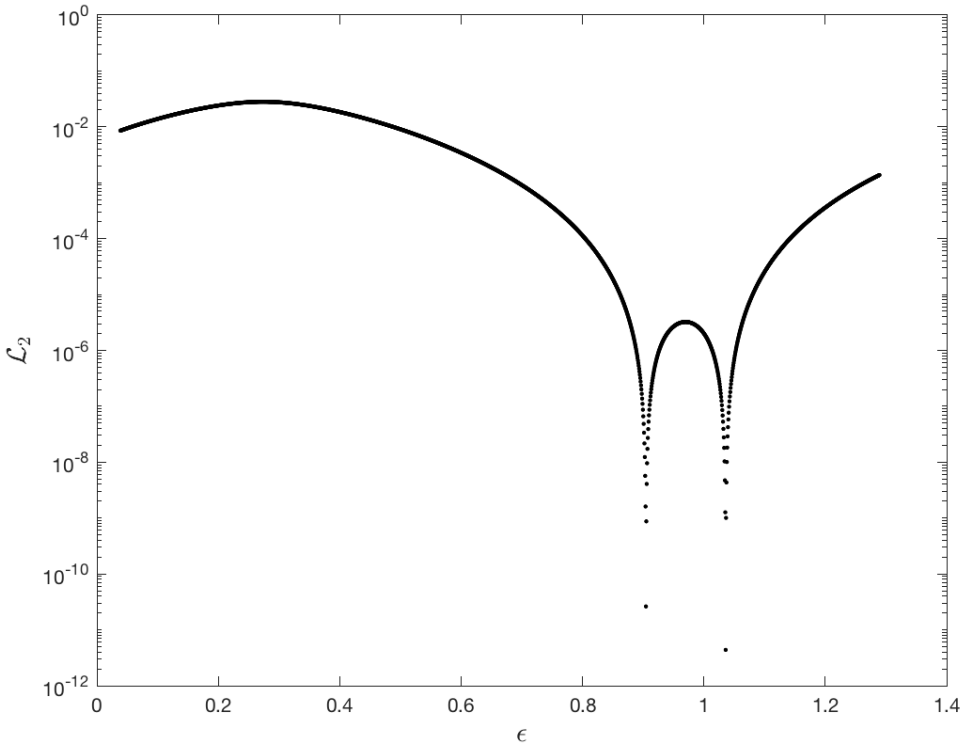


Figure 4.2: Example of the geometry problem of pointwise entry in the eigenvalue structure. The Abscissa values report the eigenvalue structure for a single eigenvalue as a point wise entry is changes over the range ϵ.

In order to apply Theorem 2, there are two things to be calculated, namely the maximum value of $f(H)$ *and the value of Hessian of* $-\log f$ *at the maximum. Maximizing* $f(H)$ *is equivalent to minimizing*

$$\phi(H) = tr\left(\Sigma^{-1} H L H^{'}\right) \tag{4.35}$$

and it is straightforward matter to show that for all $H \in O(m)$*,*

$$tr\left(\Sigma^{-1} H L H^{'}\right) \geq \sum_{i=1}^{m} \frac{l_i}{\lambda_i} \tag{4.36}$$

with equality if and only if H *is one of the* 2^m *matrices of the form*

$$\begin{bmatrix} \pm 1 & & 0 \\ & \ddots & \\ 0 & & \pm 1 \end{bmatrix} \tag{4.37}$$

The function $f(H)$ *thus has maximum of* $\exp\left[-\frac{1}{2}\sum_{i=1}^{m}(l_i/\lambda_i)\right]$ *at each of the* 2^m *matrices. The next step is to split* $O(m)$ *up into* 2^m *disjoint pieces, each containing exactly one of the matrices, and to recognize that the asymptotic behavior of each of the resulting integrals is the same. Hence for large* n*,*

$$I(n) \sim 2^m \int_{N(I_m)} [f(H)]^n (H' dH) \tag{4.38}$$

where $N(I_m)$ *is a neighborhood of the identity matrix* I_m *on the orthogonal manifold* $O(m)$*. Because the determinant of a matrix is a continuous function of the elements of the matrix we can assume that* $N(I_m)$ *contains only proper orthogonal matrices* H*. The next step involves calculating the Hessian of -logf, evaluated at* $H = I_m$*. This involves differentiating logf twice with respect to the elements of* H*. Any proper orthogonal* $m \times m$ *matrix* H *can be expressed as*

$$H = \exp(U) \equiv I_m + U + \frac{1}{2}U^2 + \frac{1}{3!}U^3 + \dots \tag{4.39}$$

where U *is an* $m \times m$ *skew-symmetric matrix. The* $\frac{1}{2}m(m-1)$ *elements of* U *provide a parametrization of* H*. The mapping* $H \to U$ *is a mapping from* $O^+(m) \to \mathbb{R}^{m(m-1)/2}$*, where* $O^+(m)$ *is the subgroup of* $O(m)$ *consisting of proper orthogonal matrices. The image of* $O^+(m)$ *under this mapping is a bounded subset*

of $\mathbb{R}^{m(m-1)/2}$. *The Jacobian of this transformation is given by*

$$(H'dH) = \bigwedge_{i<j}^{m} h_j' dh_i = \left[1 + O\left(u_{ij}^2\right)\right] \bigwedge_{i<j}^{m} du_{ij} \tag{4.40}$$

where $O(u_{ij}^r)$ *denotes terms in the* u_{ij} *which are at least of order* r. *Under the transformation* $H = \exp(U)$, $N(I_m)$ *is mapped into a neighborhood of* $U = 0$, *say,* $N^*(U = 0)$, *so that, to give*

$$I(n) \sim 2^m \int_{N^*(U=0)} [f(\exp(U))]^n (1 + \textit{higher-order terms in } U) \prod_{i<j}^{m} du_{ij} \tag{4.41}$$

Putting

$$\psi(H) = logf(H) = -\frac{1}{2} tr\left(\Sigma^{-1} H L H'\right) = -\frac{1}{2} \sum_{i,j=1}^{m} \frac{l_j}{\lambda_i} h_{ij}^2 \tag{4.42}$$

to calculate the Hessian, note that

$$-\frac{\partial^2 \psi}{\partial u_{\alpha\beta}^2} = \sum_{i,j=1}^{m} \frac{l_j}{\lambda_i} \frac{\partial^2 h_{ij}}{\partial u_{\alpha\beta}^2} + \sum_{i,j=1}^{m} \frac{l_j}{\lambda_i} \left(\frac{\partial h_{ij}}{\partial u_{\alpha\beta}}\right)^2 \tag{4.43}$$

and

$$-\frac{\partial^2 \psi}{\partial u_{\alpha\beta} \partial u_{pq}} = \sum_{i,j=1}^{m} \frac{l_j}{\lambda_i} \frac{\partial^2 h_{ij}}{\partial u_{\alpha\beta} \partial u_{pq}} + \sum_{i,j=1}^{m} \frac{l_j}{\lambda_i} \frac{\partial h_{ij}}{\partial u_{\alpha\beta}} \frac{\partial h_{ij}}{\partial u_{pq}} \tag{4.44}$$

so that in order to find the Hessian of Δ *of -logf* $= -\psi$, *it is necessary to differentiate the elements of* $H = \exp(U)$ *at most twice and set* $U = 0$. *Thus to calculate* Δ *we can use*

$$H = U + \frac{1}{2} U^2 \tag{4.45}$$

It is then a simple matter to show that, at $U = 0$,

$$-\frac{\partial^2 \psi}{\partial u_{\alpha\beta}^2} = c_{\alpha\beta} \equiv \frac{(l_\alpha - l_\beta)(\lambda_\alpha - \lambda_\beta)}{\lambda_\alpha \lambda_\beta} \tag{4.46}$$

and

$$\frac{\partial^2 \psi}{\partial u_{\alpha\beta} \partial u_{pq}} = \frac{\partial \psi}{\partial u_{\alpha\beta}} = 0 \tag{4.47}$$

so that, at $U = 0$, *the Hessian is*

$$\Delta = det \begin{bmatrix} c_{12} & & 0 \\ & \ddots & \\ 0 & & c_{m-1,m} \end{bmatrix} = \prod_{i<j}^{m} c_{ij} \tag{4.48}$$

Hence applying Theorem 2 with $p = \frac{1}{2}m(m-1)$ *shows that, for large* n,

$$I(n) \sim 2^m \exp\left(-\frac{1}{2} n \sum_{i=1}^{m} \frac{l_i}{\lambda_i}\right) \prod_{i<j}^{m} \left(\frac{2\pi}{n c_{ij}}\right)^{1/2} \tag{4.49}$$

If $H_1 \in V_{k,m}$, *the Stiefel manifold of* $m \times k$ *matrices with orthonormal columns and w choose any* $m \times (m-k)$ *matrix* H_2 *so that* $H = [H_1 : H_2] \in O(m)$ *then the unnormalized invariant measure on* $V_{k,m}$ *is*

$$(H_1^{'} dH_1) \equiv \bigwedge_{i=1}^{k} \bigwedge_{j=i+1}^{m} h_j^{'} dh_i \tag{4.50}$$

where $H = [h_1, \ldots, h_k : h_{k+1}, \ldots, h_m]$

$$\int_{V_{k,m}} (H_1^{'} dH_1) = \frac{2^k \pi^{km/2}}{\Gamma_k\left(\frac{1}{2}m\right)} \tag{4.51}$$

Given a function $f(H)$ of an $m \times m$ orthogonal matrix makes it possible to first integrate over the last $m - k$ columns of H, the first k columns being fixed, and then to integrate over these k columns.

The end of Theorem 2 proof.

The next theorem is the central result for the partitioning of the eigenvalues such that the basis within the analogue matrix are orthogonal, hence the resulting eigenvalues are valid roots.

Theorem 3. *Lemma [Muirhead, 1982].*

$$\int_{O(m)} f(H_1, H_2)(H^{'}dH) = \int_{H_1 \in V_{k,m}} \int_{K \in O(m-k)} f(H_1, GK)(K^{'}dK)(H_1^{'}dH_1) \tag{4.52}$$

where $H = [H_1 : H_2]$, H_1 *is* $m \times k$ *and* $G = G(H_1)$ *is any* $m \times (m-k)$ *matrix with orthonormal columns orthogonal to those of* H_1 *(so that* $GG^{'} = I_m - H_1 H_1^{'}$*). For the fixed* H_1*, the manifold* k_2*, say, spanned by the columns of* H_2 *can be generated by orthogonal transformations of any fixed matrix* G *chosen so that* $[H_1 : G]$ *is orthogonal; that is, any* $H_2 \in k_2$ *can be written as* $H_2 = GK$*, and as* H_2 *runs over* k_2*,* K *runs over* $O(m-k)$*, and the relationship is one-to-one. Writing*

$$H = [H_1 : H_2] = [h_1 \dots h_k : h_{k+1} \dots h_m] \tag{4.53}$$

and

$$K = [k_1 \dots k_{m-k}] \tag{4.54}$$

giving

$$dh_{k+j} = G dk_j \ (j = 1, \dots, m-k) \tag{4.55}$$

for fixed G*. Now*

$$(H^{'}dH) \equiv \bigwedge_{i<j}^{m} h_j^{'} dh_i$$

$$
\begin{aligned}
&= \bigwedge_{i<j}^{k} h_j' dh_i \bigwedge_{j=1}^{m-k} \bigwedge_{i=1}^{k} h_{k+j}' dh_i \bigwedge_{i<j}^{m-k} h_{k+j}' dh_{k+i} \\
&= \bigwedge_{i<j}^{k} h_j' dh_i \bigwedge_{j=1}^{m-k} \bigwedge_{i=1}^{k} k_j' G' dh_i \bigwedge_{i<j}^{m-k} dk_j' G' G dk_i \\
&= \bigwedge_{i<j}^{k} h_j' dh_i \bigwedge_{j=1}^{m-k} \bigwedge_{i=1}^{k} k_j' G' dh_i \bigwedge_{i<j}^{m-k} k_j' dk_i \\
&= (H_1' dH_1)(K' dK)
\end{aligned}
$$

This transformation of the measure $(H' dH)$ *is to be interpreted as: first integrate over* K *for fixed* H_1*, and then integrate over* H_1*. The end of Theorem* 3*.*

Next we need to combine the spectral result from Theorems 1 to 4 which establish the functional properties of the latent roots with the statistical object of the sample covariance matrix. Aït-Sahalia and Xiu [2018] also uses this spectral matrix result and applies it to a more general class of integrated covariance estimators.

Theorem 4. *[Muirhead, 1982].*

If $\Sigma = diag(\lambda_1, \ldots, \lambda_k, \lambda, \ldots, \lambda)$*, where*

$$\lambda_1 > \cdots > \lambda_k > \lambda \tag{4.56}$$

and the smallest root λ *is of multiplicity* $m-k$ *and* $L = diag(l_1, \ldots, l_m)$*, where* $l_1 > \cdots > l_m > 0$*, then, for large* n*,*

$${}_0F_0^{(m)}\left(-\frac{1}{2}nL, \Sigma^{-1}\right) \sim \frac{\Gamma_m\left(\frac{1}{2}m\right)}{2^m \pi^{m^2/2}} \exp\left(-\frac{1}{2}n \sum_{i=}^{k} \frac{l_i}{\lambda_i}\right) \exp\left(-\frac{n}{2\lambda} \sum_{i=k+1}^{m} l_i\right)$$

$$\cdot \prod_{i<j}^{k} \left(\frac{2\pi}{nc_{ij}} \right)^{1/2} \prod_{i=1}^{k} \prod_{j=k+1}^{m} \left(\frac{2\pi}{nd_{ij}} \right)^{1/2}$$

where

$$c_{ij} = \frac{(l_i - l_j)(\lambda_i - \lambda_i)}{\lambda_i \lambda_j} \ (i, j = 1, \ldots, k) \tag{4.57}$$

and

$$d_{ij} = \frac{(l_i - l_j)(\lambda_i - \lambda)}{\lambda \lambda_i} \ (i = 1, \ldots, k; j = k+1, \ldots, m) \tag{4.58}$$

The end of Theorem 4.

Proof. Proof of Theorem 5 The proof is similar to that of Theorem 3 but more sophisticated with the fact hat Σ has a multiple root. First, as in the proof of Theorem 3, write

$${}_0F_0^{(m)} \left(-\frac{1}{2} nL, \Sigma^{-1} \right) = \frac{\Gamma_m \left(\frac{1}{2} m \right)}{2^m \pi^{m^2/2}} I(n) \tag{4.59}$$

where

$$I(n) = \int_{O(m)} etr \left(-\frac{1}{2} n \Sigma^{-1} H L H^{'} \right) (H^{'} dH) \tag{4.60}$$

Now partition Σ and H as

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \lambda I_{m-k} \end{bmatrix}, \ \Sigma_1 = diag(\lambda_1, \ldots, \lambda_k) \tag{4.61}$$

and $H = [H_1 : H_2]$, where H_1 is $m \times k$. Then

$$tr(\Sigma^{-1} H^{'} L H)$$

$$
\begin{aligned}
&= tr(\Sigma_1^{-1} H_1' L H_1) + tr(\lambda^{-1} H_2' L H_2) \\
&= tr\left[(\Sigma_1^{-1} - \lambda^{-1} I_k) H_1' L H_1\right] + tr(\lambda^{-1} L)
\end{aligned}
$$

where I have used

$$
tr(\lambda^{-1} H_2' L H_2) = tr(\lambda^{-1} L H_2 H_2') \tag{4.62}
$$

and the fact that $H_2 H_2' = I - H_1 H_1'$. Hence

$$
I(n) = \exp\left(-\frac{n}{2\lambda}\sum_{i=1}^{m} l_i\right) \int_{O(m)} etr\left[-\frac{1}{2} n(\Sigma_1^{-1} - \lambda^{-1} I) H_1' L H_1\right] (H' dH) \tag{4.63}
$$

Applying Lemma 4 to this last integral gives

$$
\begin{aligned}
I(n) = &\exp\left(-\frac{n}{2\lambda}\sum_{i=1}^{m} l_i\right) \\
&\int_{H_1 \in V_{k,m}} \int_{K \in O(m-k)} etr\left[-\frac{1}{2} n(\Sigma_1^{-1} - \lambda^{-1} I) H_1' L H_1\right] (K' dK)(H_1' dH_1)
\end{aligned}
$$

The integrand here is not a function of K, and using the result in Corollary 2.1.16 in page 71 in Muirhead [1982] we can integrate with respect to K to give

$$
I(n) = \frac{2^{m-k} \pi^{(m-k)^2/2}}{\Gamma_{m-k}\left[\frac{1}{2}(m-k)\right]} \exp\left(-\frac{n}{2\lambda}\sum_{i=1}^{m} l_i\right) J(n) \tag{4.64}
$$

where

$$
J(n) = \int_{V_{k,m}} etr\left[-\frac{1}{2} n(\Sigma_1^{-1} - \lambda^{-1} I) H_1' L H_1\right] (H_1' dH_1) \tag{4.65}
$$

The integral $J(n)$ is of the form

$$J(n) = \int_{V_{k,m}} [f(H_1)]^n (H_1' dH_1) \tag{4.66}$$

where

$$f(H_1) = etr\left[\frac{1}{2} n(\lambda^{-1} I - \Sigma_1^{-1}) H_1' L H_1\right] \tag{4.67}$$

so that in order to apply Theorem 1 to find the asymptotic behaviour of $J(n)$ it is necessary to find the maximum value of $f(H_1)$ and the Hessian of -log$f$ at the maximum. Maximizing f is equivalent to maximizing

$$\phi(H_1) = tr\left[(\lambda^{-1} I - \Sigma_1^{-1}) H_1' L H_1\right] \tag{4.68}$$

and it follows for all $H_1 \in V_{k,m}$,

$$\phi(H_1) \leq \frac{1}{\lambda} \sum_{i=1}^{k} l_i - \sum_{i=1}^{k} \frac{l_i}{\lambda_i} \tag{4.69}$$

with equality if and only if H_1 is one of the 2^k matrices of the form

$$\begin{bmatrix} \pm 1 & & 0 \\ & \ddots & \\ 0 & & \pm 1 \\ \cdots & \cdots & \cdots \\ & 0 & \end{bmatrix} \quad (m \times k) \tag{4.70}$$

it follows that

$$J(n) \sim 2^k \int N \left(\begin{bmatrix} I_k \\ 0 \end{bmatrix} \right) [f(H_1)]^n (H_1' dH_1) \tag{4.71}$$

where $N(I_k, \ldots, 0)$ denotes a neighborhood of the matrix $(I_k, \ldots, 0)$. Now let $[H_1 : -]$ be an $m \times m$ orthogonal matrix whose first k columns are H_1. In the neighborhood above a parametrization of H_1 is given by

$$[H_1 : -] = \exp \left(\begin{bmatrix} U_{11} & U_{12} \\ -U_{12}' & 0 \end{bmatrix} \right) \tag{4.72}$$

where U_{11} is a $k \times k$ skew-symmetric matrix and U_{12} is $k \times (m-k)$. The Jacobian of this transformation is given by

$$(H_1' dH_1) = (1 + O(u_{ij}^2))(dU_{11})(dU_{12}) \tag{4.73}$$

and the projection of the vector $N((I_k, \ldots, 0)')$ under this transformation is a neighborhood, say, N^*, of $U_{11} = 0$, $U_{12} = 0$. Hence

$$J(n) \sim 2^k \int_{N^*} f^n \left[1 + O(u_{ij}^2)\right] (dU_{11})(dU_{12}) \tag{4.74}$$

To calculate the Hessian Δ of $-\log f$, put

$$\begin{aligned} \psi = \log f & \\ &= \frac{1}{2} tr \left[(\lambda^{-1} I - \Sigma_1^{-1}) H_1' L H_1 \right] \\ &= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{m} (\lambda^{-1} - \lambda_i^{-1}) l_j h_{ij}^2 \end{aligned}$$

Substitute for the h_{ji} is in terms of the elements of U_{11} and U_{12}, and evaluate $\Delta = det(-\partial^2\psi/\partial u_{ij}\partial u_{pq})$ at $U_{11} = 0$ and $U_{12} = 0$. An application of Theorem 2 then gives the asymptotic behavior of $J(n)$ for large n as

$$J(n) \sim 2^k \exp\left(\frac{n}{2\lambda}\sum_{i=1}^{k} l_i - \frac{1}{2}n\sum_{i=1}^{k}\frac{l_i}{\lambda_i}\right)\prod_{i<j}^{k}\left(\frac{2\pi}{nc_{ij}}\right)^{1/2}\prod_{i=1}^{k}\prod_{j=k+1}^{m}\left(\frac{2\pi}{nd_{ij}}\right)^{1/2}$$

Substituting this for $J(n)$

$$\frac{\Gamma_m\left(\frac{1}{2}m\right)}{\Gamma_{m-k}\left[\frac{1}{2}(m-k)\right]} = \Gamma_k\left(\frac{1}{2}m\right)\pi^{-k(k-m)/2} \tag{4.75}$$

The precise meaning of Theorem 5 is that, given $\varepsilon > 0$, there exists $n_0 \equiv n_0(\varepsilon, \Sigma, L)$ such that

$$\left|\frac{{}_0F_0^{(m)}\left(-\frac{1}{2}nL, \Sigma\right)}{h(n, L, \Sigma)} - 1\right| < \varepsilon \text{ for all } n \geq n_0 \tag{4.76}$$

■

We now have a 'Delta' method interpretation of the roots and hence the eigenvalues, Theorem 6 now provides the statistical properties of the covariance matrix and roots.

Theorem 5. *Muirhead [1982] The stochastic properties of the covariance matrix, under normality: Let $l_1, \ldots, l_m$ be the latent roots of the sample covariance matrix S formed from a sample of size $N = n+1 (n \geq m)$ from the $N_m(\mu, \Sigma)$ distribution, and suppose the latent roots $\lambda_1, \ldots, \lambda_m$ of Σ satisfy*

$$\lambda_1 > \cdots > \lambda_k > \lambda_{k+1} = \cdots = \lambda_m \ (= \lambda > 0) \tag{4.77}$$

Then for large n an asymptotic representation for the joint density function of $l_1, \ldots, l_m$ is

$$K_1 \prod_{i=1}^{k} \left[l_i^{(n-m-1)/2} \exp\left(-\frac{n l_i}{2\lambda_i} \right) \right] \tag{4.78}$$

$$\prod_{i<j}^{k} \left(\frac{l_i - l_j}{\lambda_i - \lambda_j} \right)^{1/2} \cdot \prod_{i=1}^{k} \prod_{j=k+1}^{m} \left(\frac{l_i - l_j}{\lambda_i - \lambda} \right)^{1/2} \tag{4.79}$$

$$\cdot \prod_{i=k+1}^{m} \left[l_i^{(n-m-1)/2} \exp\left(-\frac{n l_i}{2\lambda_i} \right) \right] \prod_{k+1, i<j}^{m} (l_i - l_j) \ (l_1 > \cdots > l_m > 0) \tag{4.80}$$

where

$$K_1$$

$$= \frac{\left(\frac{1}{2}n\right)^{mn/2 - k(2m-k-1/4)} \pi^{m^2/2 - k(k+1)/4} \Gamma_k\left(\frac{1}{2}m\right)}{\Gamma_m\left(\frac{1}{2}n\right) \Gamma_m\left(\frac{1}{2}m\right)} \tag{4.81}$$

$$\cdot \prod_{i=1}^{k} \lambda_i^{-(n-m+1)/2} \lambda^{-(m-k)(n-k)/2} \tag{4.82}$$

The end of Theorem 5.

4.6.1 Bias in Eigenvalue Estimation From the Sample Covariance Matrix

We have seen an example illustrated in the simulation in Figure 4.1 that the roots of the sample covariance S are biased relative to those of the underlying covariance matrix Σ. Indeed, the derivation in Theorem 6 illustrates the source of the bias, in (4.82) where the sample ratio is clearly $O(n^{1/2})$ plus an adjustment factor. For multivariate normal data, Theorem 6 provides a detailed strategy

for (1) identifying the magnitude of bias under a variety of specifications (hence permitting the derivation of pivotal statistics for tests on the underlying structure of the eigenvalues,) and hence (2) provids a strategy for writing down a Neyman-Pearson style test for the number of bounded eigenvalues. In the following two Corollaries Muirhead [1982] precisely derives the bias under normality.

Corollary 1 From Muirhead [1982], suppose that the latent roots of Σ. For large n an asymptotic representation for the conditional density function of $l_{k+1}, \ldots, l_m$, the $q = m - k$ smallest roots of S, given the k largest roots $l_1, \ldots, l_k$, is proportional to

$$\prod_{i=1}^{k} \prod_{j=k+1}^{m} (l_i - l_j)^{1/2} . \prod_{i=k+1}^{m} \left[l_i^{(n-k-q-1)/2} \exp\left(-\frac{n l_i}{2\lambda} \right) \right] \prod_{k+1, i<j}^{m} (l_i - l_j) \tag{4.83}$$

Note that this asymptotic conditional density function does not depend on $\lambda_1, \ldots, \lambda_k$, the k largest roots of Σ.

Corollary 2 Suppose the latent roots of Σ satisfy

$$\lambda_1 > \cdots > \lambda_k > \lambda_{k+1} = \cdots = \lambda_m \ (= \lambda > 0) \tag{4.84}$$

and put

$$x_i = \left(\frac{n}{2} \right)^{1/2} \left(\frac{l_i - \lambda_i}{\lambda_i} \right) \ (i = 1, \ldots, m) \tag{4.85}$$

Then the limiting joint density function of $x_1, \ldots, x_m$ as $n \to \infty$ is

$$\prod_{i=1}^{k} \phi(x_i) \frac{\pi^{q(q-1)/4}}{2^{q/2} \Gamma_q \left(\frac{1}{2} q \right)} \exp\left(-\frac{1}{2} \sum_{j=k+1}^{m} x_j^2 \right) \prod_{k+1, i<j}^{m} (x_i - x_j) \tag{4.86}$$

where $q = m - k$ and $\phi(.)$ denotes the standard normal density function.

It is interesting to look at the maximum likelihood estimates of the population latent roots obtained from the marginal distribution of the sample roots (rather than from the original normally distributed sample). The part of the joint density function of $l_1, \ldots, l_m$ involving the population roots is

$$L^* = \prod_{i=1}^{m} \lambda_i^{-n/2} \; {}_0F_0^{(m)} \left(-\frac{1}{2} nL, \Sigma^{-1} \right) \tag{4.87}$$

which we will call the marginal likelihood function. When the population roots are all distinct ($i.e., l_1 > \cdots > l_m > 0$), Theorem 3 can be used to approximate this for large n, giving

$$L^* \approx K.L_1 L_2 \tag{4.88}$$

where

$$L_1 = \prod_{i=1}^{m} \left[\lambda_i^{-n/2} \exp\left(-\frac{n}{2} \frac{l_i}{\lambda_i} \right) \right] \tag{4.89}$$

$$L_2 = \prod_{i<j}^{m} \left(\frac{\lambda_i \lambda_j}{\lambda_i - \lambda_j} \right)^{1/2} \tag{4.90}$$

and K is a constant (depending on $n, l_1, \ldots, l_m$, but not on $\lambda_1 \ldots \lambda_m$ and hence irrelevant for likelihood purposes). The values of the λ_i, which maximize L_1 are

$$\tilde{\lambda}_i = l_i \; (i = 1, \ldots, m) \tag{4.91}$$

It is easy to show that the values of the λ_i which maximize $L_1 L_2$ are

$$\hat{\lambda}_i = l_i - \frac{1}{n} l_i \sum_{j=1, j \neq i}^{m} \frac{l_j}{l_i - l_j} + O(n^{-2}) \ (i = 1, \ldots, m) \tag{4.92}$$

These estimates utilizes information from other sample roots, adjacent ones of course having the most effect. It follows easily that

$$E(\hat{\lambda}_i) = \lambda_i + O(n^{-2}) \ (i = 1, \ldots, m) \tag{4.93}$$

so that their bias terms are of order n^{-2}.

4.6.2 The distribution of distinct roots

The last piece of underlying theory is the centrepiece of the results from Anderson [1963]. This result holds for multivariate normal distributions, but we will show is very sensitive to this assumption even when the integrated covariance matrix can be unbiasedly estimated.

Theorem 6. *Anderson [1963], suppose that the latent roots of* Σ *are* $\lambda_1 \geq \cdots \geq \lambda_m > 0$, *and let* $h_1 \ldots h_m$ *be the corresponding normalized eigenvectors. Let* $q_1, \ldots, q_m$ *be the normalized eigenvectors of the sample covariance matrix* S *corresponding to the latent roots* $l_1 > \cdots > l_m > 0$. *If* λ_i *is a distinct root then, as* $n \to \infty$, $n^{1/2}(q_i - h_i)$ *has a limiting* m*-variate normal distribution with mean* 0 *and covariance matrix*

$$\Gamma = \lambda_i \sum_{j=1, j \neq i}^{m} \frac{\lambda_j}{(\lambda_i - \lambda_j)^2} h_j h_j' \tag{4.94}$$

and is asymptotically independent of l_i

The end of Theorem 6.

4.6.3 Classical inference problems on Latent Roots

The final step is to write down a series of test statistics to determine the structure of the latent roots. Our objective is to determine how many are bounded, hence implying the number of unbounded roots, which correspond to the number of uncorrelated factors implied within the data.

The following basic classic are given in Muirhead [1982] which is a formalization of the results first described in [Anderson, 1963].

The main objective is to impute a bound of the following form:

$$H_k : \lambda_{k+1} = \cdots = \lambda_m \tag{4.95}$$

for $k = 0, 1, \ldots, m-2$, where $\lambda_1 \geq \cdots \geq \lambda_m > 0$ are the latent roots of Σ. The likelihood ratio test of

$$H_0 : \lambda_1 = \cdots = \lambda_m \tag{4.96}$$

is based on the statistic

$$V_0 = \frac{\prod_{i=1}^{m} l_i}{\left(\frac{1}{m}\sum_{i=1}^{m} l_i\right)^m} \tag{4.97}$$

where $l_1 > \cdots > l_m$ are the latent roots of the sample covariance matrix S, and a test of asymptotic size α is to reject H_0 if

$$-\left(n - \frac{2m^2 + m + 2}{6m}\right) logV_0 > c\left(\alpha; \frac{1}{2}(m+2)(m-1)\right) \tag{4.98}$$

where $c(\alpha; r)$ denotes the upper $100\alpha\%$ point of the $\chi^2_{1/2(q+2)(q-1)}$ distribution.

Theorem 7. *adapted from Anderson [1963] and Muirhead [1982] Given a sample of size N from the $N_m(\mu, \Sigma)$ distribution, the likelihood ratio statistic for testing the null hypothesis*

$$H_k : \lambda_{k+1} = \cdots = \lambda_m \ (= \lambda, unknown) \tag{4.99}$$

is $\wedge_k = V_k^{N/2}$, where

$$V_k \equiv \frac{\prod_{i=k+1}^{m} l_i}{\left(\frac{1}{m-k}\sum_{i=k+1}^{m} l_i\right)^{m-k}} \tag{4.100}$$

When H_k is true, the maximum value of the likelihood function is

$$\exp\left(-\frac{n}{2\hat{\lambda}_i}\sum_{i=k+1}^{m} l_i - \frac{1}{2}n\sum_{i=1}^{k}\frac{l_i}{\hat{\lambda}_i}\right)\left(\prod_{i=1}^{k}\hat{\lambda}_i^{-N/2}\right)\hat{\lambda}_i^{-N(m-k)/2} \tag{4.101}$$

where $n = N - 1$, and

$$\hat{\lambda}_i = \frac{n}{N}l_i (i = 1, \ldots, k), \hat{\lambda} = \frac{1}{m-k}\frac{n}{N}\sum_{i=k+1}^{m} l_i \tag{4.102}$$

are the maximum l likelihood estimates of the λ_i and λ under H_k. Substituting for these gives the maximum of the likelihood function under H_k as

$$\max_{H_k} L(\mu, \Sigma) = \left(\frac{N}{n}\right)^{mN/2}\left(\prod_{i=1}^{k} l_i^{-N/2}\right)\left(\frac{1}{m-k}\sum_{i=k+1}^{m} l_i\right)^{-N(m-k)/2} e^{-mN/2} \tag{4.103}$$

The end of Theorem 7

When μ and Σ are unrestricted the maximum value of the likelihood function given by

$$\underset{\mu,\Sigma}{max} L(\mu, \Sigma) = \left(\frac{N}{n}\right)^{mN/2} \left(\prod_{i=1}^{m} l_i^{-N/2}\right) e^{-mN/2} \tag{4.104}$$

so that the likelihood ratio statistic for testing H_k is given by

$$\wedge_k = \frac{\underset{H_k}{max} L(\mu, \Sigma)}{\underset{\mu,\Sigma}{max} L(\mu, \Sigma)} = \left[\frac{\prod_{i=k+1}^{m} l_i}{\left(\frac{1}{m-k}\sum_{i=k+1}^{m} l_i\right)^{m-k}}\right]^{N/2} = V_k^{N/2} \tag{4.105}$$

Rejecting H_k for small values of $\wedge_k$ is equivalent to rejecting H_k for small values of V_k, and the proof is complete. The end of Theorem 7 proof.

Let us now turn to the asymptotic distribution of the statistic V_k when the null hypothesis H_k is true. It is convenient to put $q = m - k$ and

$$\bar{l}_q = \frac{1}{q}\sum_{i=k+1}^{m} l_i \tag{4.106}$$

the average of the smallest q latent roots of S, so that

$$V_k = \frac{\prod_{i=k+1}^{m} l_i}{\bar{l}_q^q} \tag{4.107}$$

The general theory of likelihood ratio tests shows that, as $n \to \infty$, the asymptotic distribution of $-n \log V_k$ is $\chi^2_{(q+2)(q-1)/2}$ when H_k is true. An improvement over $-n \log V_k$ is the statistic

$$-\left(n - k - \frac{2q^2 + q + 2}{6q}\right) \log V_k \ (q = m - k) \tag{4.108}$$

We noted the discussion following Corollary 2 that when H_k is true the asymptotic conditional density function of $l_{k+1}, \ldots, l_m$, the q smallest latent roots of S, given the k largest roots $l_1, \ldots, l_k$, does not depend on $\lambda_1, \ldots, \lambda_k$, the k largest roots of Σ. In a test of H_k these k largest roots are nuisance parameters; the essential idea of James is that the effects of these nuisance parameters can be eliminated, at least asymptotically, by testing H_k using this conditional distribution. If we put

$$u_i = \frac{l_i}{\bar{l}_q} \ (i = k+1, \ldots, m) \tag{4.109}$$

in the asymptotic conditional density function of $l_{k+1}, \ldots, l_m$, given $l_1, \ldots, l_k$ in Corollary 2, then the asymptotic density function of $u_{k+1}, \ldots, u_{m-1}$, conditional on $l_1, \ldots, l_k, \bar{l}_q$, follows easily as

$$K_2 \prod_{i=1}^{k} \prod_{j=k+1}^{m} (r_i - u_j)^{1/2} \prod_{i=k+1}^{m} u_i^{(n-k-q-1)/2} \prod_{k+1, i<j}^{m} (u_i - u_j) \tag{4.110}$$

where $r_i = l_i/\bar{l}_q$ for $i = 1, \ldots, k$, and K_2 is a constant. Note that $\sum_{i=k+1}^{m} u_i = q$ and that

$$V_k = \prod_{i=k+1}^{m} \left(\frac{l_i}{\bar{l}_q} \right) = \prod_{i=k+1}^{m} u_i \tag{4.111}$$

Put $T_k = -\log V_k$ so that the limiting distribution of nT_k is $\chi^2_{(q+2)(q-1)/2}$ when H_k is true. The appropriate multiplies of T_k can be obtained by finding its expected value. For notational convenience, let $\mathbb{E}_c$ denote expectation taken with respect to the conditional distribution of $u_{k+1}, \ldots, u_{m-1}$ given $l_1, \ldots, l_k, \bar{l}_q$ and let $\mathbb{E}_N$ denote

expectation taken with respect with respect to the "null" distribution

$$K_3 \prod_{i=k+1}^{m} u_i^{(n-k-q-1)/2} \prod_{k+1,i<j}^{m} (u_i - u_j) \tag{4.112}$$

where K_3 is constant, obtained by ignoring the linkage factor

$$\prod_{i=1}^{k} \prod_{j=k+1}^{m} (r_i - u_j)^{1/2} \tag{4.113}$$

4.6.4 Determining whether the statistic is a pivot

For bootstrap analysis of any test statistic the only requirement is that the distribution under the null should not be affected by any model specific features, see MacKinnon [2002] and MacKinnon [2006] for the philosophical discussion on inference on bootstrap. That is, we do not need to know the true model structure to identify the distribution of the test statistic under the null. For instance, we have an unbiased estimate of the mean of a random variable and its variance.

If the variable is IID normally distributed, then the distribution of the mean divided by the squareroot of the variance under the asumption that the mean is zero can be determined by drawing random samples of equivalent length data from a distribution with a zero mean and variance identical to the sample variance and constructing the same ratio. Inference is then based on the 100α percentiles as needed. Furthermore, as the series is presumed to be normal, be can sample with replacement the series, subtract the resampled mean and compute the resampled variance, to give a distribution of the test statistic. In both cases the test statistic is generated under the null of the mean being zero. In both cases

the resampling is normally distributed, in the first case by construction and in the second case using the presumed properties of the actual data.

Unfortunately, bootstrapping a z-score is one of the few tests where derivation of the pivot under the null is simple enough to illustrate without recourse to the underlying stochastic properties. Theorems 9 and 10 are re-derived from Muirhead [1982] in our notation, to show the pivot features indeed, the intention of Muirhead [1982] was to derive the Neyman-Pearson form of the statistic, but this is useful as it serves the same purpose for the bootstrap.

Theorem 8. *Muirhead [1982] ,when the null hypothesis H_k is true, the limiting distribution, as $n \to \infty$, of the statistic*

$$P_k = -\left[n - k - \frac{2q^2 + q + 2}{6q} + \sum_{i=1}^{k} \frac{\bar{l}_q^2}{(l_i - \bar{l}_q)^2}\right] logV_k \tag{4.114}$$

is $\chi^2_{(q+2)(q-1)/2}$, and

$$\mathbb{E}_c(P_k) = \frac{1}{2}(q+2)(q-1) + O(n^{-2}) \tag{4.115}$$

Proof. Proof of Theorem 8 [Muirhead, 1982]

$$\begin{aligned}
&\mathbb{E}_c(T_k) \\
&\quad = \mathbb{E}_c\left(-ln \prod_{i=k+1}^{m} u_i\right) \\
&\quad = \mathbb{E}_c\left[-\frac{\partial}{\partial h}\left(\prod_{i=k+1}^{m} u_i^h\right)_{h=0}\right] \\
&\quad = -\frac{\partial}{\partial h}\left[\mathbb{E}_c\left(\prod_{i=k+1}^{m} u_i^h\right)\right]_{h=0}
\end{aligned}$$

$$= -\frac{\partial}{\partial h}[\mathbb{E}_c(e^{-hT_k})]_{h=0}$$

We can exchange the order of differentiation, see Apostol [1969, Volume 2, Chapter 14] and integration because in a neighborhood of $h = 0$

$$h^{-1}\left(1 - \prod_{i=k+1}^{m} u_i^h\right) \leq 2\sum_{i=k+1}^{m} u_i = 2q \tag{4.116}$$

Hence, in order to find $\mathbb{E}_c(T_k)$ we will first obtain

$$\mathbb{E}_c(e^{-hT_k}) = \mathbb{E}_c\left(\prod_{i=k+1}^{m} u_i^h\right) \tag{4.117}$$

This can obviously be done by finding

$$\mathbb{E}_N\left[\prod_{i=1}^{k}\prod_{j=k+1}^{m}(r_i - u_j)^{1/2}.\exp(-hT_k)\right] \tag{4.118}$$

Now, when H_k is true,

$$1 - u_j = O_p(n^{-1/2}) \tag{4.119}$$

so that

$$\begin{aligned}(r_i - u_j)^{1/2} &= (r_i - 1)^{1/2}\left(1 + \frac{1-u_j}{r_i - 1}\right)^{1/2} \\ &= (r_i - 1)^{1/2}\left[1 + \frac{1}{2}\frac{(1-u_j)}{(r_i-1)} - \frac{1}{8}\frac{(1-u_j)^2}{(r_i-1)^2} + O_p(n^{-3/2})\right]\end{aligned}$$

Since $\sum_{j=k+1}^{m}(1-u_j)=0$, we get

$$\begin{aligned}
&\prod_{i=1}^{k}\prod_{j=k+1}^{m}(r_i-u_j)^{1/2} \\
&\qquad=\prod_{i=1}^{k}(r_i-1)^{q/2}\left[1+\frac{1}{2(r_i-1)^2}\sum_{k+1,j<p}^{m}(1-u_j)(1-u_p)+O_p(n^{-3/2})\right] \\
&\qquad=\prod_{i=1}^{k}(r_i-1)^{q/2}.\left\{1+\frac{1}{2}\alpha\left[\sum_{k+1,i<j}^{m}u_iu_j-\binom{q}{2}\right]+O_p(n^{-3/2})\right\}
\end{aligned}$$

where $q=m-k$ and

$$\alpha=\sum_{i=1}^{k}\frac{1}{(r_i-1)^2}=\sum_{i=1}^{k}\frac{\bar{l}_q^2}{\left(l_i-\bar{l}_q\right)^2} \tag{4.120}$$

and finally, to get the expectation of quadratic variation we need to decompose the following expectation:

$$\mathbb{E}_N\left[\left(\sum_{k+1,i<j}^{m}u_iu_j\right)\exp(-hT_k)\right]=\mathbb{E}_N\left[\left(\sum_{k+1,i<j}^{m}u_iu_j\right)\prod_{i=k+1}^{m}u_i^h\right] \tag{4.121}$$

This problem is addressed in the following lemma using the seminal result of Balakrishnan [2006, Volume 2, Chapter 4].

■

Theorem 9.

$$\mathbb{E}_N\left[\left(\sum_{k+1,i<j}^{m}u_iu_j\right)\exp(-hT_k)\right]=\binom{q}{2}\left(\frac{n-k-1+2h}{n-k+2/q+2h}\right)\mathbb{E}_0(h) \tag{4.122}$$

where

$$\mathbb{E}_0(h) \equiv \mathbb{E}_N[\exp(-hT_k)] = \mathbb{E}_N\left(\prod_{i=k+1}^{m} u_i^h\right) \tag{4.123}$$

Proof. Proof of Theorem 9 Muirhead [1982], since $u_i = l_i/\bar{l}_q$ for $i = k+1, \ldots, m$, it follows that

$$\left(\sum_{k+1,i<j}^{m} l_i l_j\right) \prod_{i=k+1}^{m} l_i^h = \bar{l}_q^{qh+2} \left(\sum_{k+1,i<j}^{m} u_i u_j\right) \prod_{i=k+1}^{m} u_i^h \tag{4.124}$$

The null distribution of $l_{k+1}, \ldots, l_m$ is the same as the distribution of the latent roots of a $q \times q$ covariance matrix S such that $(n-k)S$ has the $\mathcal{W}_q(n-k, \lambda I_q)$ distribution, so that I will regard $l_{k+1}, \ldots, l_m$ as the latent root of the sample covariance matrix S. All posterior expectations involving l_i for $i = k+1, \ldots, m$ are taken with respect to this distribution. Put $n^{'} = n-k$; then $(n^{'}/\lambda)S$ is $\mathcal{W}_q(n^{'}, I_q)$, $(n^{'}/\lambda)trS = (n^{'}/\lambda)q\bar{l}_q$ is $\chi^2_{n^{'}q}$, from which it follows easily that

$$E(\bar{l}_q^r) = \left(\frac{1}{2}\frac{n^{'}}{\lambda}q\right)^{-r} \left(\frac{1}{2}n^{'}q\right)_r \tag{4.125}$$

where $(x)_r = x(x+1)\ldots(x+r-1)$. Moreover, $\bar{l}_q$ is independent of u_i, $i = k+1, \ldots, m$, and hence

$$\frac{\mathbb{E}_N\left[\left(\sum_{k+1,i<j}^{m} u_i u_j\right)\prod_{i=k+1}^{m} u_i^h\right]}{\mathbb{E}_N\left(\prod_{i=k+1}^{m} u_i^h\right)} = \frac{E\left[\left(\sum_{k+1,i<j}^{m} l_i l_j\right)\prod_{i=k+1}^{m} l_i^h\right] E\left(\bar{l}_q^{qh}\right)}{E\left(\bar{l}_q^{qh+2}\right) E\left(\prod_{i=k+1}^{m} l_i^h\right)} \tag{4.126}$$

where we have used the fact that

$$\prod_{i=k+1}^{m} l_i = \bar{l}_q^q \prod_{i=k+1}^{m} u_i \tag{4.127}$$

Now

$$\prod_{i=k+1}^{m} l_i = \det S \tag{4.128}$$

and

$$\prod_{k+1,i<j}^{m} l_i l_j = r_2(S) \tag{4.129}$$

the accumulation of second-order (2×2) principal minors of S. Since the principal minors all give the same expectation, I need only to find the expectation involving the first one to find the bound

$$\Delta = det \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix} \tag{4.130}$$

We can then multiply by

$$\begin{pmatrix} q \\ 2 \end{pmatrix} \tag{4.131}$$

the number of them, Put $(n^{'}/\lambda)S = T^{'}T$, where $T = (t_{ij})$ is a $q \times q$ upper-triangular matrix and by construction t_{11}^2 are independent $\chi^2_{n^{'}-i+1}$ random variables $(i = 1, \ldots, q)$, from which it is easy to verify that

$$\begin{aligned} E\left(\prod_{i=k+1}^{m} l_i^h\right) & \\ &= E[(detS)^h] \\ &= \left(\frac{n^{'}}{\lambda}\right)^{-qh} E\left(\prod_{i=1}^{q} t_{11}^{2h}\right) \\ &= \left(\frac{n^{'}}{\lambda}\right)^{-qh} \prod_{i=1}^{q} \left(\frac{1}{2}(n^{'}-i+1)\right)_h \end{aligned}$$

and

$$
\begin{aligned}
&E\left[\left(\sum_{k+1,i<j}^{m} l_i l_j\right)\prod_{i=k+1}^{m} l_i^h\right] \\
&\qquad = \binom{q}{2} E[\Delta(detS)^h] \\
&\qquad = \binom{q}{2} E\left[t_{11}^2 t_{22}^2 \prod_{i=1}^{q} t_{11}^{2h}\right]\left(\frac{n^{'}}{\lambda}\right)^{-(qh+2)} \\
&\qquad = \binom{q}{2}\left(\frac{1}{2}\frac{n^{'}}{\lambda}\right)^{-(qh+2)}\left(\frac{1}{2}n^{'}\right)_{h+1}\left(\frac{1}{2}(n^{'}-1)\right)_{h+1} \\
&\qquad \cdot \prod_{i=3}^{q}\left(\frac{1}{2}(n^{'}-i+1)\right)_h
\end{aligned}
$$

Substitution gives

$$
\begin{aligned}
&\frac{\mathbb{E}_N\left[\left(\sum_{k+1,i<j}^{m} u_i u_j\right)\prod_{i=k+1}^{m} u_i^h\right]}{\mathbb{E}_N\left(\prod_{i=k+1}^{m} u_i^h\right)} \\
&\qquad = \frac{\binom{q}{2}\left(\frac{1}{2}n^{'}\right)_{h+1}\left(\frac{1}{2}(n^{'}-1)\right)_{h+1} q^2\left(\frac{1}{2}n^{'}q\right)_{qh}}{\left(\frac{1}{2}n^{'}\right)_h\left(\frac{1}{2}(n^{'}-1)\right)_h\left(\frac{1}{2}n^{'}q\right)_{qh+2}} \\
&\qquad = \binom{q}{2}\frac{\left(\frac{1}{2}n^{'}+h\right)\left(\frac{1}{2}n^{'}-\frac{1}{2}+h\right)q^2}{\left(\frac{1}{2}n^{'}q+qh\right)\left(\frac{1}{2}n^{'}q+qh+1\right)} \\
&\qquad = \binom{q}{2}\frac{\frac{1}{2}n^{'}-\frac{1}{2}+h}{\frac{1}{2}n^{'}+1/q+h}
\end{aligned}
$$

$$= \binom{q}{2} \frac{n^{'} - 1 + 2h}{n^{'} + 2/q + 2h}$$
$$= \binom{q}{2} \frac{n - k - 1 + 2h}{n - k + 2/q + 2h}$$

which completes the proof of the Theorem 9.

Let us now explain the proof of Theorem 8 :

$$\mathbb{E}_c(e^{-hT_k}) = \frac{\theta(h)}{\theta(0)} \tag{4.132}$$

where

$$\theta(h) = \mathbb{E}_0(h) f(h) \tag{4.133}$$

with

$$f(h) = 1 + \frac{1}{2}\alpha \binom{q}{2} \left[\frac{n - k - 1 + 2h}{n - k + \frac{2}{q} + 2h} - 1 \right] \tag{4.134}$$

Thus

$$\mathbb{E}_c(T_k) = -\frac{\partial}{\partial h} \left[\frac{\theta(h)}{\theta(0)} \right]_{h=0} = -\mathbb{E}_0^{'}(0) - \frac{f^{'}(0)}{f(0)} = -\mathbb{E}_0^{'}(0) - \frac{\alpha d}{n^2} + O(n^{-3}) \tag{4.135}$$

where $d = (q-1)(q+2)/2$ and α is the critical bound. We know from Section 8.3 in Muirhead [1982] that $[n - k - (2q^2 + q + 2)/6q]T_k$ has an asymptotic χ^2_d distribution as $n \to \infty$, and the means agree to $O(n^{-2})$ so that

$$-\mathbb{E}_0^{'}(0) = \frac{d}{n - k - (2q^2 + q + 2)/6q} + O(n^{-3}) \tag{4.136}$$

Substitution gives

$$\mathbb{E}_0(T_k) = \frac{d}{n-k-(2q^2+q+2)/6q} - \frac{\alpha d}{n^2} + O(n^{-3}) \tag{4.137}$$

from which it follows that if P_k is the statistic defined then

$$\mathbb{E}_c(P_k) = d + O(n^{-2}) \tag{4.138}$$

and the proof is complete.

■

It follows from Theorem 3 that if n is large an approximate test of size α of the null hypothesis

$$H_k : \lambda_{k+1} = \cdots = \lambda_m \tag{4.139}$$

is to reject H_k if $P_k > c(\alpha; (q+2)(q-1)/2)$, where P_k is given, $q = m - k$ and $c(\alpha; r)$ is the upper $100\alpha\%$ point of the χ^2_r distribution. An estimate of λ is provided by

$$\bar{l_q} = q^{-1} \sum_{i=k+1}^{m} l_i \tag{4.140}$$

and it is easy to show, for example, that as $n \to \infty$ the asymptotic distribution of $\left(\frac{1}{2}nq\right)^{1/2} (\bar{l_q} - \lambda)/\lambda$ is standard normal $\mathcal{N}(0,1)$. Let z_α be the upper $100\alpha\%$ point of the $\mathcal{N}(0,1)$ distribution, that is, such that $\Phi(z_\alpha) = 1 - \alpha$, where $\Phi(.)$ denotes the standard normal distribution function. Then asymptotically,

$$P\left(\left(\frac{nq}{2}\right)^{1/2}\left(\frac{\bar{l_q} - \lambda}{\lambda}\right) \geq -z_\alpha\right) = 1 - \alpha \tag{4.141}$$

which appoints to a one-sided confidence interval for λ, namely,

$$\lambda \leq \frac{\bar{l}_q}{1 - z_\alpha(2/nq)^{1/2}} \tag{4.142}$$

with asymptotic confidence coefficient $1-\alpha$. If the upper limit of this confidence interval is sufficiently small we might decide that λ is negligible and study only the first k principal components.

Even if we cannot conclude that some of the smallest latent roots of Σ are equal, it still may be possible that the variation explained by the last $q = m - k$ principal components, namely $\sum_{i=k+1}^{m} \lambda_i$, is small compared with the total variation $\sum_{i=1}^{m} \lambda_i$, in which case we might decide to study only the first k principal components. Thus it is of interest to consider the null hypothesis

$$H_k^* : \frac{\sum_{i=k+1}^{m} \lambda_i}{\sum_{i=1}^{m} \lambda_i} = h \tag{4.143}$$

where $h(0 < h < 1)$ is a number to be specified by the experimenter. This can be tested using the statistic

$$M_k \equiv \sum_{i=k+1}^{m} l_i - h \sum_{i=1}^{m} l_i = -h \sum_{i=1}^{k} l_i + (1-h) \sum_{i=k+1}^{m} l_i \tag{4.144}$$

Assuming the latent roots of Σ are distinct, Corollary 2 shows that the limiting distribution as $n \to \infty$ of

$$n^{1/2} \left[M_k + h \sum_{i=1}^{k} \lambda_i - (1-h) \sum_{i=k+1}^{m} \lambda_i \right] \tag{4.145}$$

is normal with mean 0 and variance

$$\tau^2 = 2h^2 \sum_{i=1}^{k} \lambda_i^2 + 2(1-h)^2 \sum_{i=k+1}^{m} \lambda_i^2 \tag{4.146}$$

Replacing λ_i by $l_i (i = 1, \ldots, m)$ in τ^2, this result can be used to construct an approximate test of H_k^* and to give confidence intervals for

$$\sum_{i=k+1}^{m} \lambda_i - h \sum_{i=1}^{m} \lambda_i \tag{4.147}$$

Finally, let me derive an asymptotic test for a given principal component. Let H^{**} be the null hypothesis that the vector of coefficients h_1 of the first principal components is equal to an specified $m \times 1$ vector h_1^0,i.e.,

$$H^{**} : h_1 = h_1^0,\ h_1^{0'} h_1^0 = 1 \tag{4.148}$$

Recall that h_1 is the eigenvector of Σ corresponding to the largest latent root λ_1; we will assume that λ_1 is a distinct root. A test of H^{**} can be constructed using the result of Theorem 7, namely, that if q_1 is the normalized eigenvector of the sample covariance matrix S corresponding to the largest latent root l_1 of S then the asymptotic distribution of $y = n^{1/2}(q_1 - h_1)$ is $N_m(0, \Gamma)$, where

$$\Gamma = \sum_{i=2}^{m} \frac{\lambda_1 \lambda_i}{(\lambda_1 - \lambda_i)^2} h_i h_i^{'} = H_2 B^2 H_2^{'} \tag{4.149}$$

with $H_2 = [h_2 \dots h_m]$ and

$$B^2 = \begin{bmatrix} \frac{\lambda_1\lambda_2}{(\lambda_1-\lambda_2)^2} & & & 0 \\ & \frac{\lambda_1\lambda_3}{(\lambda_1-\lambda_3)^2} & & \\ & & \ddots & \\ 0 & & & \frac{\lambda_1\lambda_m}{(\lambda_1-\lambda_m)^2} \end{bmatrix} \tag{4.150}$$

Note that the covariance matrix Γ in this asymptotic distribution is singular, as is to expected. Put $z = B^{-1}H_2^{'}y$; then the limiting distribution of z is $\mathcal{N}_{m-1}(0, I_{m-1})$, and hence the limiting distribution of $z^{'}z$ is χ^2_{m-1}. Now note that

$$z^{'}z = y^{'}H_2B^{-2}H_2^{'}y \tag{4.151}$$

and the matrix of this quadratic form in y is

$$\begin{aligned} & H_2B^{-2}H_2^{'} \\ & = H_2 \begin{bmatrix} \frac{\lambda_1}{\lambda_2} - 2 + \frac{\lambda_2}{\lambda_1} & & & 0 \\ & \frac{\lambda_1}{\lambda_3} - 2 + \frac{\lambda_3}{\lambda_1} & & \\ & & \ddots & \\ 0 & & & \frac{\lambda_1}{\lambda_m} - 2 + \frac{\lambda_m}{\lambda_1} \end{bmatrix} H_2^{'} \\ & = \lambda_1 H_2 \begin{bmatrix} \frac{1}{\lambda_2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda_m} \end{bmatrix} H_2^{'} - 2H_2H_2^{'} + \frac{1}{\lambda_1}H_2 \begin{bmatrix} \lambda_2 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix} H_2^{'} \\ & = \lambda_1 \sum_{i=2}^{m} \frac{1}{\lambda_i} h_i h_i^{'} - 2H_2H_2^{'} + \frac{1}{\lambda_1} \sum_{i=2}^{m} \lambda_i h_i h_i^{'} \end{aligned}$$

Putting $\wedge = diag(\lambda_1, \ldots, \lambda_m)$ and using

$$\Sigma = H \wedge H^{'} = \sum_{i=1}^{m} \lambda_i h_i h_i^{'} \tag{4.152}$$

$$\Sigma^{-1} = H \wedge^{-1} H^{'} = \sum_{i=1}^{m} \frac{1}{\lambda_i} h_i h_i^{'} \tag{4.153}$$

and

$$H_2 H_2^{'} = I - h_1 h_1{}' \tag{4.154}$$

becomes

$$\begin{aligned} & H_2 B^{-2} H_2^{'} \\ & \qquad = \lambda_1 \left(\Sigma^{-1} - \frac{1}{\lambda_1} j_1 h_1^{'} \right) - 2(I - h_1 h_1^{'}) + \frac{1}{\lambda_1}(\Sigma - \lambda_1 h_1 h_1^{'}) \\ & \qquad = \lambda_1 \Sigma^{-1} - 2I + \frac{1}{\lambda_1}\Sigma \end{aligned}$$

Hence the limiting distribution of

$$\begin{aligned} n(q_1 - h_1)^{'} \left(\lambda_1 \Sigma^{-1} - 2I + \frac{1}{\lambda_1}\Sigma \right) (q_1 - h_1) & \\ = n q^{'} \left(\lambda_1 \Sigma^{-1} - 2I + \frac{1}{\lambda_1}\Sigma \right) q_1 & \\ = n \left(\lambda_1 q_1^{'} \Sigma^{-1} q_1 + \frac{1}{\lambda_1} q_1^{'} \Sigma q_1 - 2 \right) & \end{aligned}$$

is χ^2_{m-1}. Since S, S^{-1}, and l_1 are consistent estimates of Σ, Σ^{-1}, and λ_1, they can be substituted for Σ, Σ^{-1}, and λ_1 without affecting the limiting distribution.

Hence, when $H^{**}: h_1 = h_1^0$ is true, the limiting distribution of

$$\begin{aligned} & W \\ & \quad = n(q_1 - h_1^0)' \left(l_1 S^{-1} - 2I + \frac{1}{l_1} S \right) (q_1 - h_1^0) \\ & \quad = n \left(l_1 h_1^{0\prime} S^{-1} h_1^0 + \frac{1}{l_1} h_1^{0\prime} S h_1^0 - 2 \right) \end{aligned}$$

is χ^2_{m-1}. It follows that a test of H^{**} of asymptotic size α is to reject H^{**} if $W > c(\alpha; m-1)$, where $c(\alpha; m-1)$ is the upper $100\alpha\%$ point of the χ^2_{m-1} distribution.

4.6.5 Sequencing

An important point to note is that 'correct' null hypothesis occurs only once within the data set. Consider that the true number of bounded eigenvalues is $q^* = m - k^*$. That is, there are k^* uncorrelated factors and there is an unknown variance σ^2 determining the variation of the cross-section of uncorrected noise. Hence the standard factor model from Chamberlain [1983] holds:

$$\Sigma = \Lambda\Lambda' + \sigma^2 I$$

The correctly sized χ^2 test is only valid when we test $k = k^*$. However, we will need to implement the test sequentially when $k < k^*$ and when $k > k^*$. Figure 4.3 provides an example of a power function for testing if the mean of a normal distribution is equal to zero.

As the true mean deviates from zero, the power of the test to correctly reject

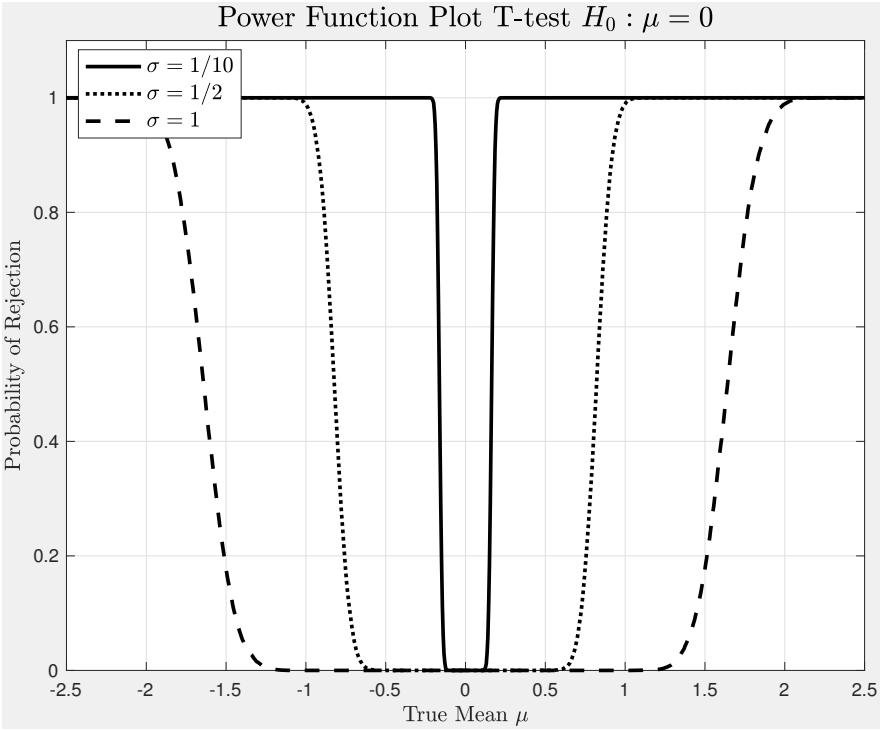


Figure 4.3: Classic Power Function Plot for a Normal distribution, with three different generating variances. Of the three functions, $\sigma = 1/10$ clearly has the highest power to correctly evaluate the null, in this case the distribution of the test statistic is independent of the actual specification.

the null that $H_0 : \mu = 0$ increases. However, in this case the true mean μ^* can vary continuously and the distribution of the sample mean is constant across all values of μ^* for a given variance.

In the case of our test the distribution of the test statistic is *only* valid at the correctly specified null. In this case the number of factors k is adjusted stepwise until we reject. One approach could be to attempt a Bonferoni style correction. However, the weighting of each $k < k^*$ depends on the value of k^*.It could pre-test using an information criteria like the one specified in Aït-Sahalia and Xiu [2018], however, this overly complicates the problem. We propose two steps: first, bias correction of the eigenvalues from the presumed DGP. Second, determine the distribution of the test statistic under the null sequentially. Power function analysis then allows us to evaluate the potential for either stopping the stepwise progression of k too early (too few factors) or stopping too late (too many). Hence, the test can approached in a conservative or liberal manner with regard to the the number of factors. For instance, 95% certainty that there are at least k^* factors or 95% certainty that there are a minimum of k^* factors. For asset pricing we normally go with the former interpretation, but it is very simple to re-task the analysis to account for the latter.

The following chapter will look at various statistical methods that correct the extremal eigenvalues and then specify a bootstrap procedure to test specifically for the underlying factor structure.

Chapter 5

Bootstrap Corrections of Extremal Eigenvalues

5.1 Introduction

The previous chapter established the classical limit theory for the eigenvalue structure of integrated covariance in our context. Whilst the standard theory developed in the 1950s and 1960s by T W Anderson see Anderson [1959] and Anderson [1963] and primarily and further developed by Muirhead [1982]and Bartlett in their 1982 book provide the necessary background for simple distributions, the application to Levy processes has not been fully considered. In this chapter we will briefly review some new theoretical results outlined in Aït-Sahalia and Xiu [2018] primarily and Onatski [2010] amongst others.

The approach will be as follows: first establish the information criteria to eigenvalue structure (and hence the factor structure) then develop a diagnostic

test using bootstrap to determine the size and power of the test. This progresses the state-of-the-art to yield an exact test which can have rejection rates compared to those under the theoretical distribution given the data generating process is indeed under the correct null.

5.2 Thresholding and information criteria

Aït-Sahalia and Xiu [2018] proposes an information criteria, which achieves an unbiased extrema at the correct number of factors under a Brownian Semi Martingale $\mathcal{BSM}$ type process. Alternative approaches, such as Onatski [2010] ask the question: how many factors are required to explain at least $100\alpha\%$ of the quadratic variation of the set of variables. Both approaches have attractive features and drawbacks. First, Aït-Sahalia and Xiu [2018] is mathematically elegant, in the sense that an asymptotic theory is determined for the eigenvalue structure. However, the penalty function is essentially arbitrary and their Monte-Carlo simulations establish unbiasedness, but not whether the distribution is a pivot, which it is not as a ratio in the Neyman-Pearson sense test cannot be constructed from the extremal distributions, see Aït-Sahalia and Xiu [2018] for details. Similarly, threshold tests such as Onatski [2010] rely on an a-priori assumption on the ratio of factors to noise. Recall that under the Chamberlain [1983] factor model structure we presume that: $\Sigma = \Lambda\Lambda' + \sigma^2 I$. Now the size of σ compared to det $(\Lambda\Lambda')$, which is unobserved, determines this ratio. It is perfectly reasonable to presume that $\sigma/\det(\Lambda\Lambda')$ is quite large, hence setting $100\alpha\%$ too big will result in a large number of redundant factors being chosen. Similarly, if $\sigma/(\det \Lambda\Lambda')$ is very small,

choosing a low $100\alpha\%$ will result in insufficient factors being selected. Hence, a-priori thresholding with such a nuisance parameter is not so useful when trying to identify the specific number of bounded eigenvalues, i.e. those which are equal, as this number is not necessarily correlated with the fraction of quadratic variation explained by the k-th marginal latent factor. Our approach *directly* tests for boundedness in a Neyman-Pearson set up. That it determines a uniformly powerful test to determine the k^* factors under the correctly specified null.

5.3 Empirically determining the bias of Extremal Eigenvalues

We will now illustrate that the asymptotic distribution of l_i is biased under even the most benign conditions and design a bootstrap to correct this bias and subsequently recover the test statistic. Using Monte-Carlo simulations We will then compute the power functions for the asymptotic test for the number of components, versus the classical short sample and our bootstrap corrected method. Classical theory says that for a set of eigenvalues $l_1 > l_2, \ldots, l_m$ for the sample realized covariance matrix S, can be expressed as a likelihood ratio statistic of the following form:

$$\bar{l_q} = \frac{1}{q} \sum_{i=k+1}^{m} l_i \tag{5.1}$$

the average of the smallest q= m-k latent roots of S, so that

$$V_k = \frac{\prod_{i=k+1}^{m} l_i}{\bar{l}_q^q} \tag{5.2}$$

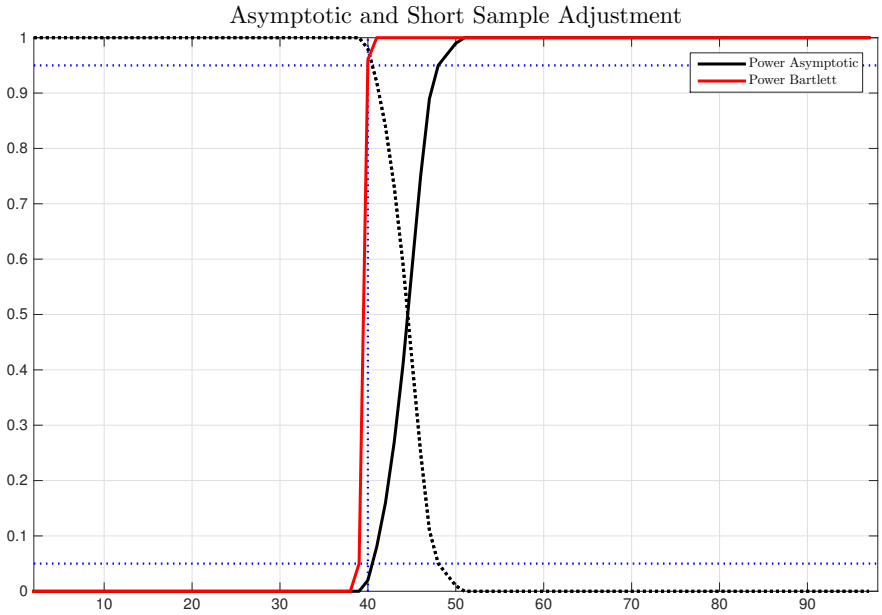


Figure 5.1: Power Function Comparison.
Notes: The power function for the iid case, comparing the Asymptotic and Bartlett sample corrected tests.

when the smallest $q = m - k$ eigenvalues are identical we can think of the covariance matrix can be decomposed into the following form:

$$\Sigma = \Lambda\Lambda' + \sigma I$$

where Λ is an $m \times q$ matrix and I is an m identity matrix. Hence we have the reduced rank matrix $\Lambda\Lambda'$ of factors and the full rank noise σI.

The disadvantage of this identification strategy is that it assumes that each asset in the list has the same noise, we will show that the bootstrap can go some way to alleviating this. For instance, if $\boldsymbol{\sigma}$ is a vector of independent trading noise for each asset, then :

$$\Sigma = \Lambda\Lambda' + \text{diag}(\boldsymbol{\sigma})$$

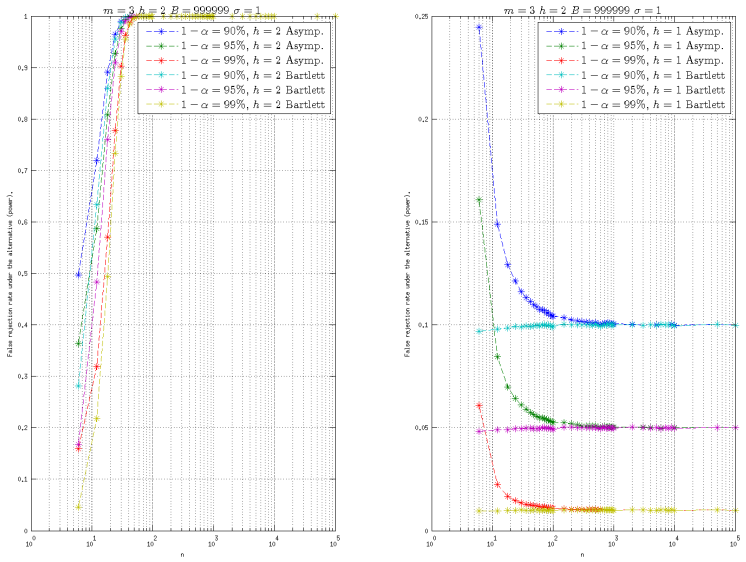


Figure 5.2: Power Function Small Cross Section: Sample Size Comparison under the null, low noise.
Testing under the null for confidence levels 90%,95%,99% asymptotically in Bartlett method.

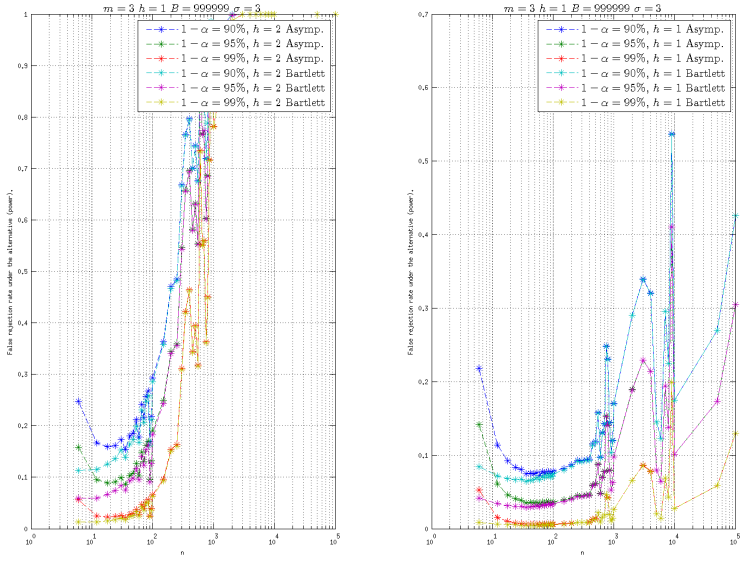


Figure 5.3: Power Function Small Cross Section: Sample Size Comparison under the null, high noise.
Testing under the null for confidence levels 90%,95%,99% in asymptotically in Bartlett method.

Unfortunately, a general theory does not exist (as the range of distributions for σ_i is very large), but some examples of when the approach will work and when it breaks down can be illustrated. In Figure 5.4 we have conducted a small simulation starting with a fixed covariance matrix Σ, we have then computed the sample estimator S from a sample of $N = 500$ observations and drawn 2,000 replications to compute the Empirical Distribution Function (EDF) of the sample error on the latent roots of S, $l_i - \lambda_i$, where as before, λ_i is the i-th eigenvalue of Σ. As we noted before in the statistical theory section the asymptotic theory provides an attempt at a small sample correction using the Bartlett method Bartlett [1963]

$$\mathscr{V}_k = -\left(n - k - \tfrac{2q^2+q+2}{6q}\right)\log[V_k] \sim \chi^2(\tfrac{1}{2}(q+2)(q-1)) \tag{5.3}$$

An obvious approach would be to directly bootstrap $\mathscr{V}_k$ to attempt to recover the empirical distribution of the test statistic and then return the appropriate critical value. However, it is well known that a χ^2 test is neither an asymptotically pivotal statistic nor a sample pivot see MacKinnon [1992] for illustration. Hence, an alternative is to bootstrap l_i directly and correct for the sample bias, then recompute the critical statistic by simulation under the null sequentially (usefully, this only needs to be done once). Our two main simulations are for (1) a large cross-section of assets ($m = 100$) and (2) a small cross section sampled at a wide range of frequencies $m = 20$).

Figure 5.1 plots the power function for the asymptotic and Bartlett test statistics. Note, the major that issue, whilst the Bartlett test has power (the function

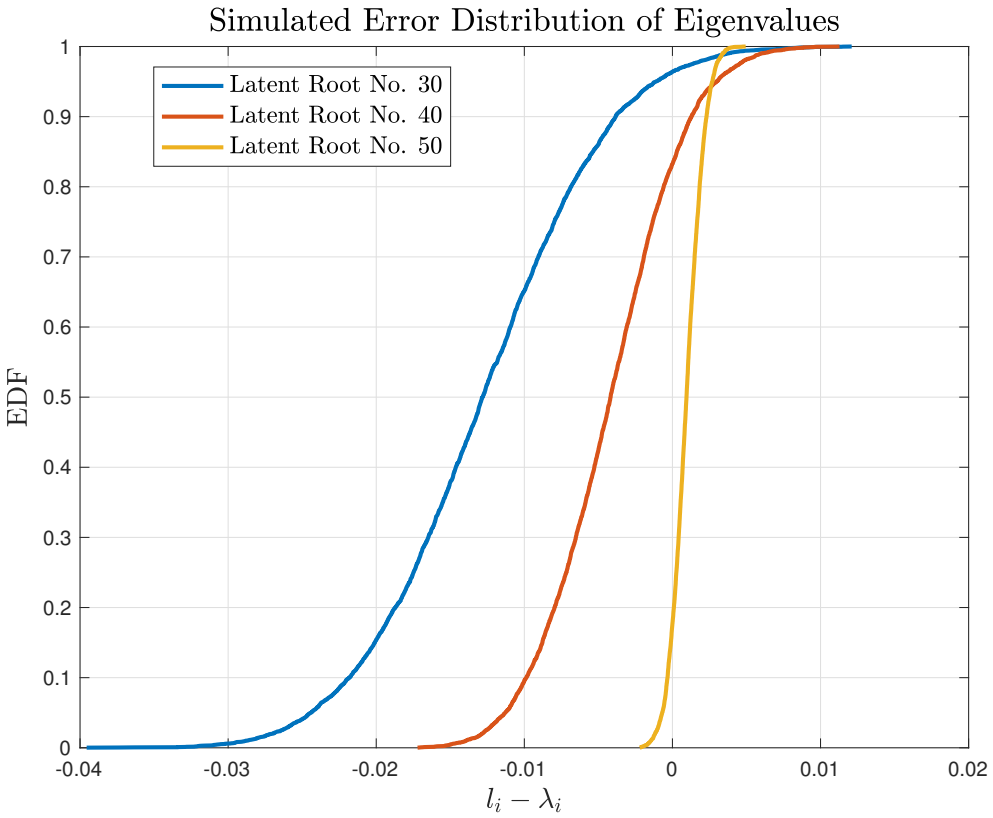


Figure 5.4: Empirical Distribution Function of the Error of Latent Roots. *Notes:* The Empirical Distribution Function (EDF) of the error on latent roots of the sample covariance matrix, from 2,000 simulations of a 100-variate ($m = 100$) generated using 40 factors $k = 40$ with an average signal to noise ratio of $9:1$ (average variance of the factors to the noise variance σ^2). The plot is for eigenvalues numbered 30, 40 and 50. I can see clearly that the probability mass of the error is not symmetrical about zero.

is very steep), it is biased (the red line crosses the 5% rejection rate at a lower number of detected factors than the correct black line). Unfortunately, whilst the rejection rate for the asymptotic test is correct at $k = 40$, the test lacks power, as the solid black line has a much gentler slope suggesting that the test will likely include many spurious factors. In this case the signal-to-noise ratio is low, at 10%. However, as this rises toward 50:50, the power drops markedly, sample size, is also critical, see Figures 5.2 and 5.3, which pivots the power function to plot another case, but varying the number of observations $n = N - 1$. In each case the power and consistency of the test break down in many common situations.

5.4 A simple bootstrap correction

Two obvious methods for constructing a consistent bootstrap test for signal rank are (1) to pre-pivot the bootstrap, so that the asymptotic distribution of $\mathscr{V}_k$ is approximately $\chi^2((q+2)(q-1)/2)$ or (2) correct the sample bias for the estimation of l_i from the sample covariance matrix S.

Several recent working papers have proposed approaches for (1) including Williams, Dovonon, and Taamouti [2017], (testing the number of factors in high frequency data) working paper hence I have chosen to look more closely at option (2), often referred to as 'bias correction'. Here, I will generate draws under the sample distribution assuming that the sample covariance matrix is unbiased for the generation of the latent roots, then recompute the empirical distribution of the bias corrected test statistic under the null directly.

Notes that there are a variety of estimators for the sample realized covariance estimator, ranging from the traditional maximum-likelihood estimator to the more complex estimators covered in Chapters 2 and 3. However, the major difficulty is in constructing the distribution of the test statistic under the null and we will explain our approach to this in the next subsection.

5.4.1 Generating samples under the null

It is worth reviewing the intuition from the classical result in Theorem 8, for a set of sorted eigenvalues of the sample covariance, approximating the standardized latent roots, we are presuming that a $q = m - k$ block of the remaining smallest eigenvalues are identical and hence, this is uncorrelated white noise with an approximately identical variance.

Let, $\boldsymbol{l} = [l_1, \ldots, l_m]'$, a vector of sorted eigenvalues (largest to smallest) and $\boldsymbol{V}$ be the equivalently sorted eigenvectors of the full rank sample covariance matrix S. Hence $S = \boldsymbol{V}\text{diag}[\boldsymbol{l}]\boldsymbol{V}^{-1}$. To generate the equivalent matrix under the null it is necessary to recondition $(\boldsymbol{l})$ and $(\boldsymbol{V})$ to the presumed true latent structure. The matrix $\boldsymbol{V}$ is assumed to have columns $\boldsymbol{v}_i$ indexed by $i \in \{1, \ldots, m\}$.

Hence we construct $\boldsymbol{l}_k^\dagger$ and $\boldsymbol{V}_k^\dagger$, where the last q$= m - k$ entries are averaged in the same manner as the standard test statistic as follows:

$$\bar{l}_q = 1/q \sum_{i=k+1}^{m} l_i, \quad \bar{\boldsymbol{v}}_q = 1/q \sum_{i=k+1}^{m} \boldsymbol{v}_i$$

we then replace the original elements $k + 1$ to m elements of $\boldsymbol{l}$ and $\boldsymbol{V}$ with their

average quantities to generate:

$$\mathbf{l}_k^\dagger = [l_1, \ldots, l_k, \bar{l}_k, \ldots, \bar{l}_k], \quad \boldsymbol{V}_k^\dagger = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k, \bar{\boldsymbol{v}}_k, \ldots, \bar{\boldsymbol{v}}_k],$$

and hence construct $S_k^\dagger = \boldsymbol{V}_k^\dagger \text{diag}[\boldsymbol{l}_k^\dagger](\boldsymbol{V}_k^\dagger)^{-1}$, which is the covariance matrix under the null hypothesis that the $q = m - k$ entries are identical.

The bootstrap-in-bootstrap then proceeds as follows:

1. Construct an unbiased estimate of the sample covariance matrix S.
2. Compute the sorted sample eigenvalues and eigenvectors $\boldsymbol{l}$ and $\boldsymbol{V}$.
3. For each step k, generate the sample equivalent under the null hypothesis

$$S_k^\dagger = \boldsymbol{V}_k^\dagger \text{diag}[\boldsymbol{l}_k^\dagger](\boldsymbol{V}_k^\dagger)^{-1}$$

4. Generate an equivalent sized block of simulated data Y^* using the covariance matrix $S_k^\dagger$ under the null hypothesis and compute a draw $S_k^{\dagger *}$ as the equivalent sample covariance matrix using the same approach used to compute S, the simplest case is that the data is multivariate normal, hence $Y^* = \bar{Y} + \boldsymbol{E}(S_k^{\dagger *})^{1/2}$, where $(S_k^{\dagger *})^{1/2}$ is the Cholesky factor of $S_k^{\dagger *}$ and $\bar{Y}$ is a matrix with rows replicating the unbiased mean vector of Y.
5. Compute the test statistic $\mathscr{V}_k^{\dagger *}$ from the matrix $S_k^{\dagger *}$ estimated from the simulated sample in the standard fashion.
6. Repeat steps 2 and 3 a large number of times about 999.

7. Sort the recovered $\mathscr{V}_k^{\dagger *}$ and recover the statistical $1-\alpha$ bound of interest (for instance the 95-th percentile).

Theorem 10. *New Result: Bootstrap Consistency Theorem. When* $Y^* \sim \mathcal{N}(0, I \otimes \Sigma)$*, hence* S *is the maximum likelihood estimate of* Σ *with* $Y^* \sim \mathcal{N}(0, I \otimes S_k^{\dagger *})$ *and* $Y^* \sim \mathcal{N}(0, I \otimes S_k^{\dagger *})$ *then the asymptotic distribution of* $\mathscr{V}_k^{\dagger *}$ *is* $\chi^2[(p+2)(p-1)/2]$ *converges to the bootstrap diostribution.*

Proof. The proof follows directly from the asymptotic proof for the distribution of the LR statistic $\mathscr{V}_k$ in Theorem Lemma 9 reproduced from Anderson [1959] and Muirhead [1982], as we directly simulating the distribution of the restricted matrix under the null. The bootstrap consistency is derived from the fact that the resampled data sets Y^* are normal subsamples, hence yielding the same asymptotic distribution as Y. ■

5.4.2 Notes

Theorem 10 demonstrates the mechanistic result that the bootstrap directly simulates the covariance matrix under the null. But the key benefit of the bootstrap is when the data generating process is non-Gaussian. In this case we can extend the myriad of results on realized covariance matrices to the PCA domain.

Figure 5.5 illustrates the sample consistency of the distribution of the bootstrap under when the data is under the null and the test choosing at the correct boundary point. For small sample sizes the test is slightly conservative at the 95-th percentile in this case which is designed to approximate 5 minute data for 100 assets at five minutes for a week.

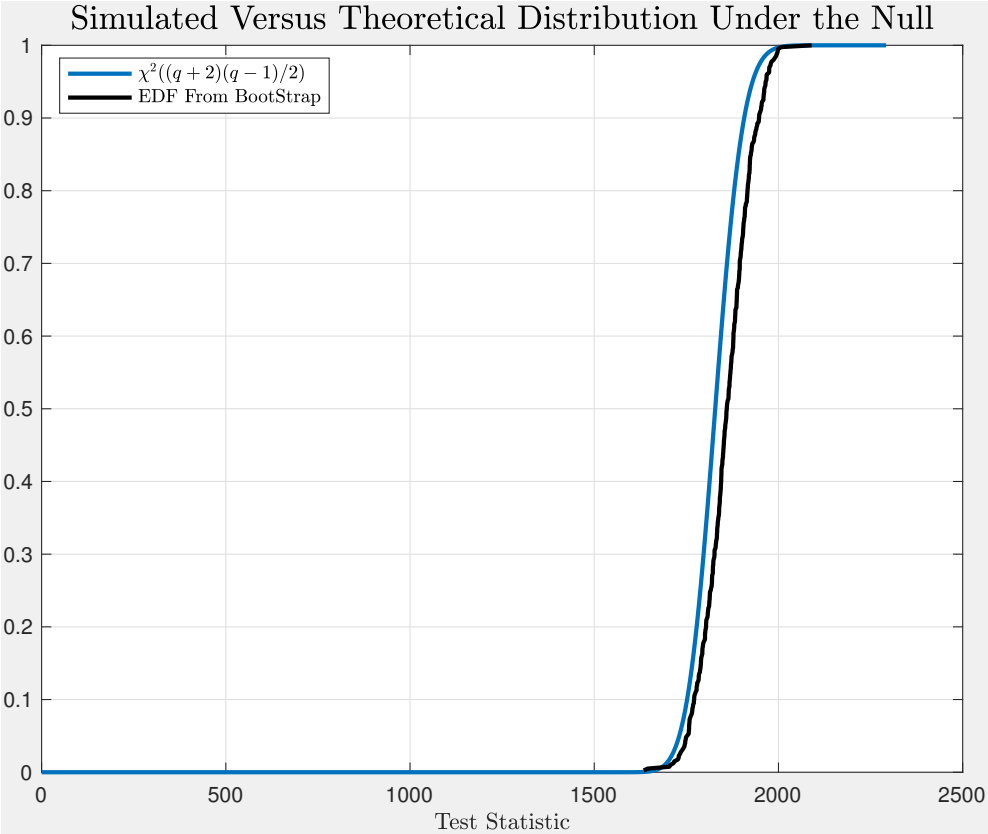


Figure 5.5: Comparison of the distribution of the PCA test statistic for data simulated under the null (in this case $m = 100$, the true number of factors is $k = 40$ and σ is set such that the average signal variance to noise ratio is 0.2. Data is IID normal. Data is generated from 500 observations, hence 5 observations per cross section.

5.5 Power function analysis of the bootstrap

Figures 5.6 to 5.15 present the power functions comparing the performance of the Bootstrap versus the classical tests under a variety of simulation conditions. It is important to note that this function deviates slightly from the classical interpretation of a power function, in the sense that we seek to reject the null for factors below the correct value for k, denoted $\bar{k}$ from the data generating process and then fail to reject for integer values of k greater than the true value under the true data generating process. Hence, we plot both the false rejection and false fail to reject frequencies as these have a countervailing interpretation for values of k above and below $\bar{k}$.

The steeper the gradient of the curve the more power the test has in discriminating between model specifications. However, to be correctly sized, the curve should pass through the points denoted by the intersection of the blue lines (5% for the power function under the null and 95% for the power function under the alternative).

The bootstrap correctly sizes and has power in almost all cases, whereas the classical tests, even with the Bartlett correction can be both under and over-sized and, when the signal to noise ratio is high, significantly lack power.

Figure 5.16 shows the first and second principal components for the cross section of S&P 500 stocks for one month of five minute data. Using the bootstrap we have constructed the 95% confidence bounds for the two components and this illustrates a common finding, that the first market factor can be estimated with a high degree of precision, but the second component is very noisy, with a huge confidence bound.

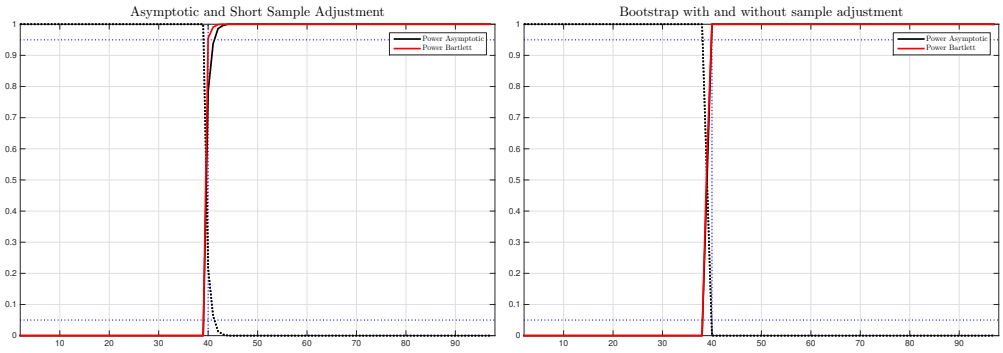


Figure 5.6: Power Function Comparison for sample size $N = 2000$, dimension $m = 100$, factor dimension $k = 40$ for the data generating process and signal to noise variance ratio of 0.2.

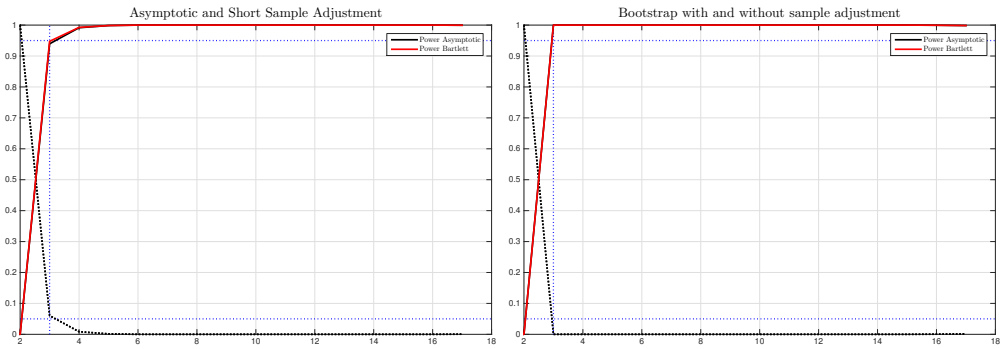


Figure 5.7: Power Function Comparison for sample size $N = 2000$, dimension $m = 20$, factor dimension $k = 3$ for the data generating process and signal to noise variance ratio of 0.01.

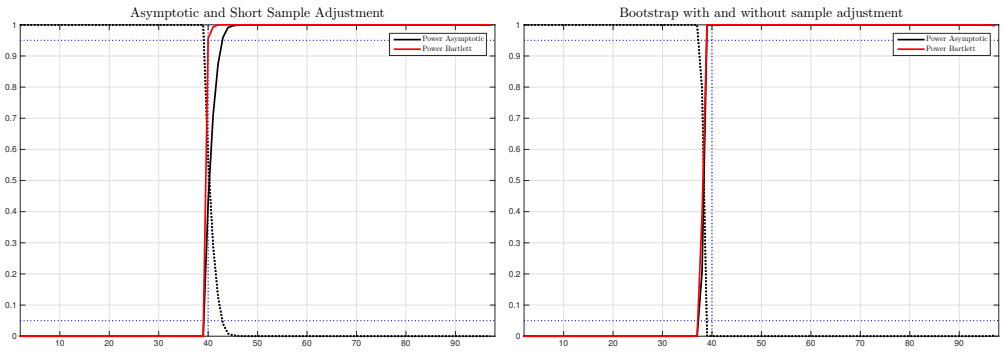


Figure 5.8: Power Function Comparison for sample size $N = 1000$, dimension $m = 100$, factor dimension $k = 40$ for the data generating process and signal to noise variance ratio of 0.2.

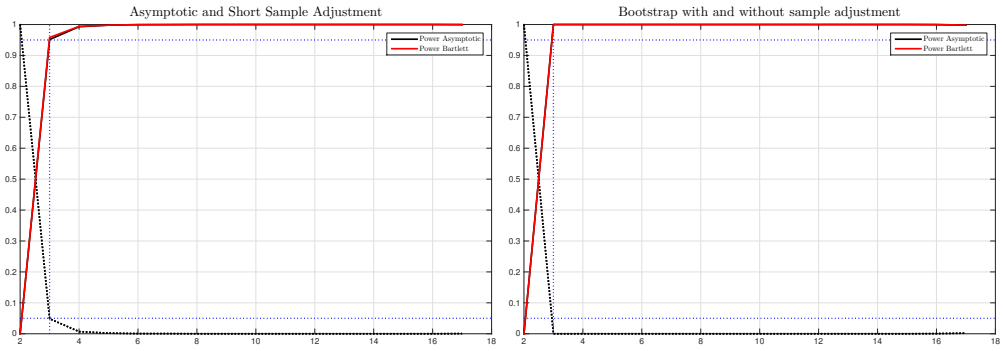


Figure 5.9: Power Function Comparison for sample size $N = 1000$, dimension $m = 20$, factor dimension $k = 3$ for the data generating process and signal to noise variance ratio of 0.01.

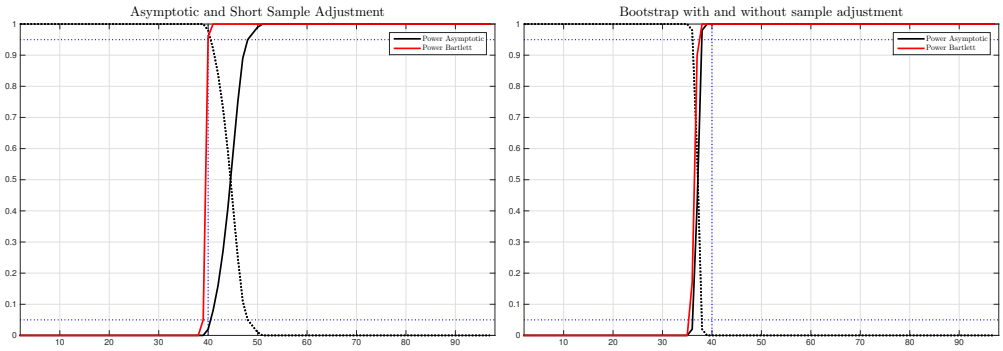


Figure 5.10: Power Function Comparison for sample size $N = 500$, dimension $m = 100$, factor dimension $k = 40$ for the data generating process and signal to noise variance ratio of 0.2.

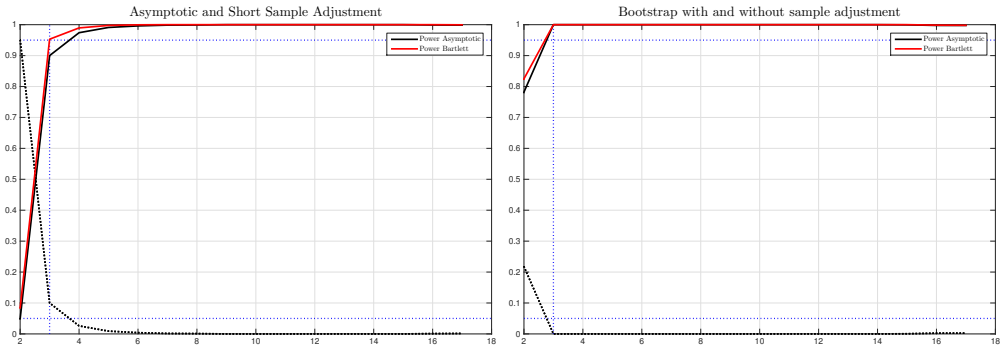


Figure 5.11: Power Function Comparison for sample size $N = 200$, dimension $m = 20$, factor dimension $k = 3$ for the data generating process and signal to noise variance ratio of 0.5.

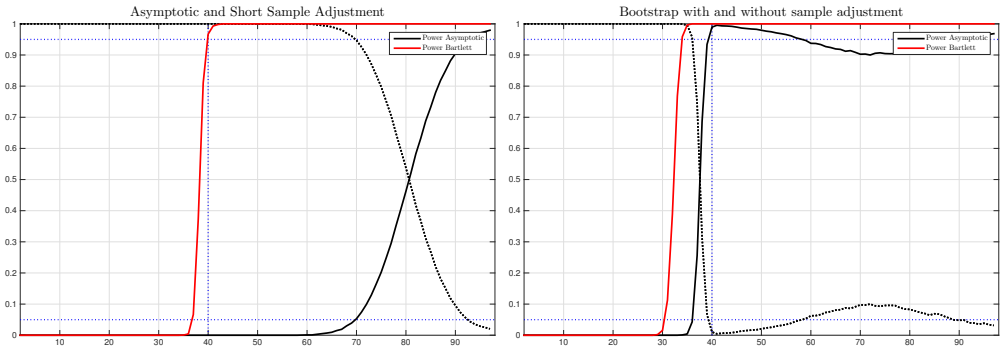


Figure 5.12: Power Function Comparison for sample size $N = 200$, dimension $m = 100$, factor dimension $k = 40$ for the data generating process and signal to noise variance ratio of 0.2.

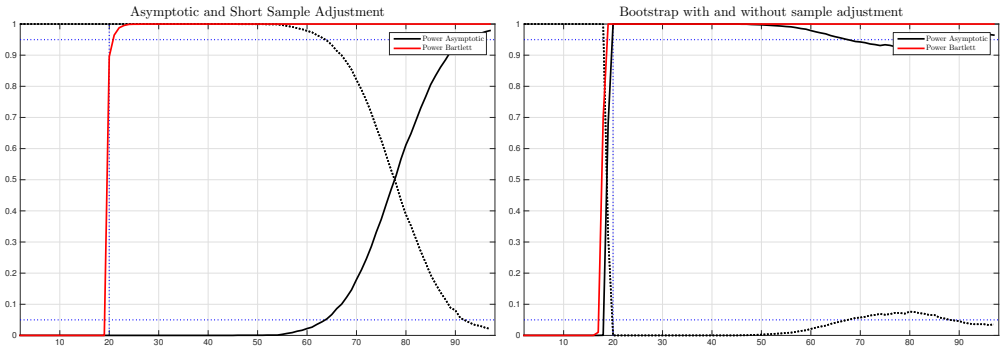


Figure 5.13: Power Function Comparison for sample size $N = 200$, dimension $m = 100$, factor dimension $k = 20$ for the data generating process and signal to noise variance ratio of 0.2.

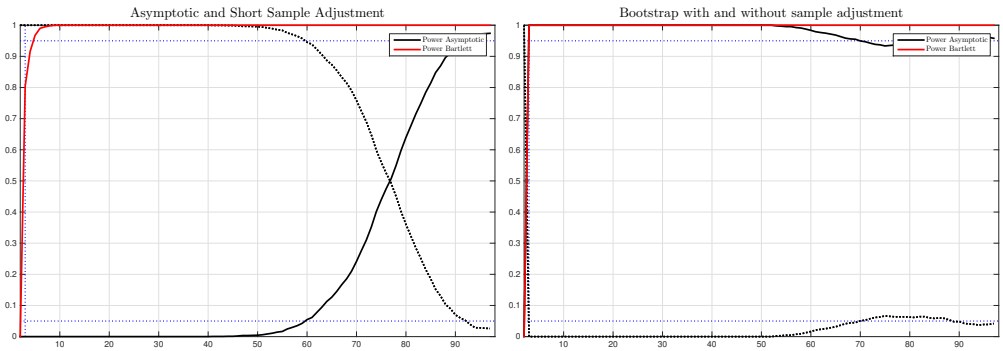


Figure 5.14: Power Function Comparison for sample size $N = 200$, dimension $m = 100$, factor dimension $k = 3$ for the data generating process and signal to noise variance ratio of 0.2.

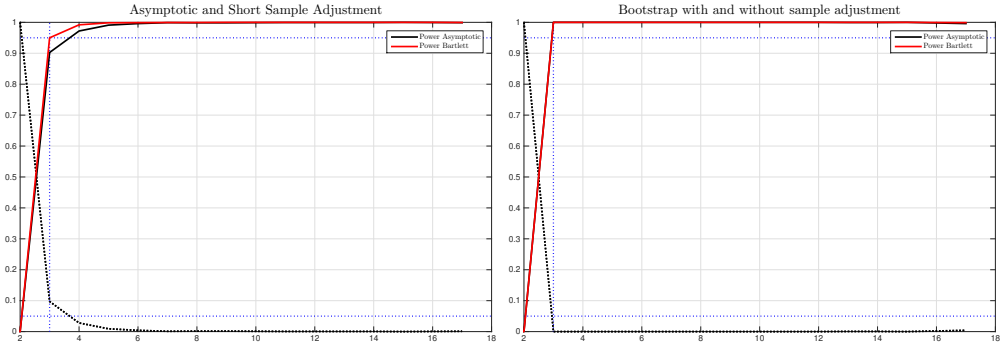


Figure 5.15: Power Function Comparison for sample size $N = 200$, dimension $m = 20$, factor dimension $k = 3$ for the data generating process and signal to noise variance ratio of 0.2.

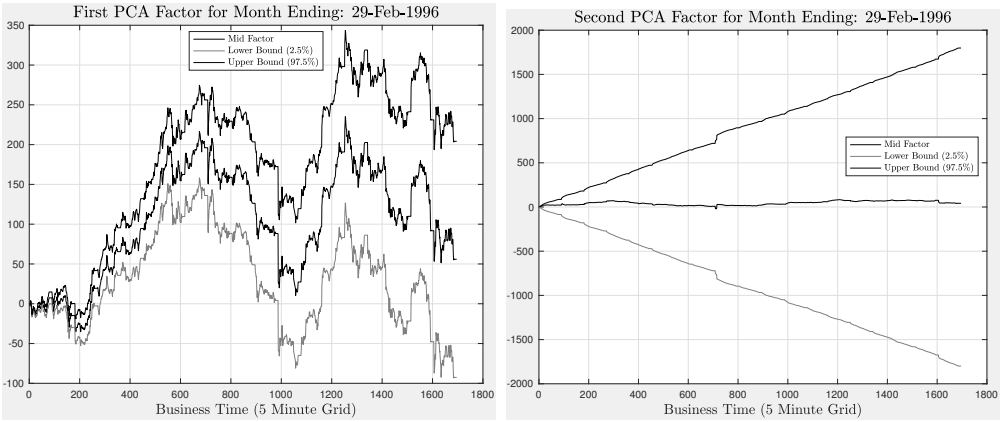


Figure 5.16: First and Second Principle Components and Confidence Bounds for the S&P 500 Cross section from 5 minute data for the business time during the month ending February 29, 1996.

5.6 Extracting the factor structure of WTI crude oil future prices

The main empirical analysis of this chapter focuses on detecting the number of factors in the term structure of WTI futures prices. The problem of empirically extracting the latent factor dynamics is quite apparent. Figures 5.17 to 5.21 present the term structure and cumulative return for a variety of days for WTI futures contracts. Figure 5.17 presents the term structure for August 10, 2011, here we see a jump in the 2.4 year tenor futures contract. The red star for each tenor represents the median price for the day and the term structure is a spline through the median contracts. The jump causes problems for any analysis as this is a clear regime shift in the futures term structure. However, clear structures remain from 6.4 to 9.4 years. The right plot presents the side-by-side time evolution of the cumulative return for each contract. The jump clearly occurs just after 12 noon. But there are many of highly volatile returns across the futures curve, with several contracts sparsely traded until *after* the sudden change in the 2.4 year contract.

The prices after 12pm represent the ex-mark to market prices. Hence we see a lot of volatility in the contracts pre and post this shut down period. The term structure in Figure 5.18 illustrates this issue quite well, several kinks can be seen in the term structure and clearly if we are trying to compute a covariance matrix, there are joint regime shifts within the variation and correlation structure over the day. Hence estimation of integrated covariance and the resulting distribution of eigenvalues will be affected by these regime shifts. Figures 5.19 and 5.21 also

have this complex volatility structure. Figures 5.19 and 5.20 illustrate slightly less noisy days, but we still observe significant variation in the pricing activity across tenors.

Thus, we can determine some interesting stylized facts. First, the number of turning points in the term structure does not necessarily reflect the number of factors within the price evolution of the data. Volatility across tenors and through the time evolution of the series is not constant. Jumps are prevalent factor in the data and need to be accounted for in both the integrated covariance estimation and in determining the distribution of the test statistic critical bounds that then allow us to infer the underlying factor structure.

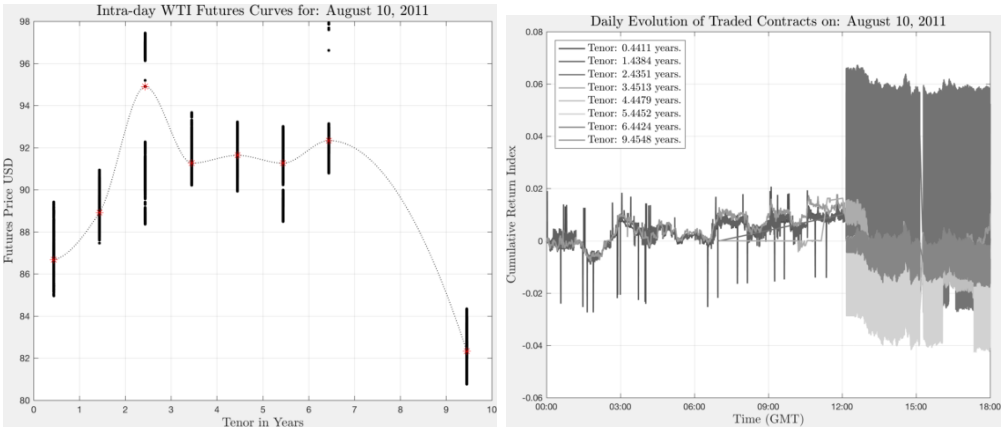


Figure 5.17: The term structure and cumulative return evolution for the WTI futures market for a single day.

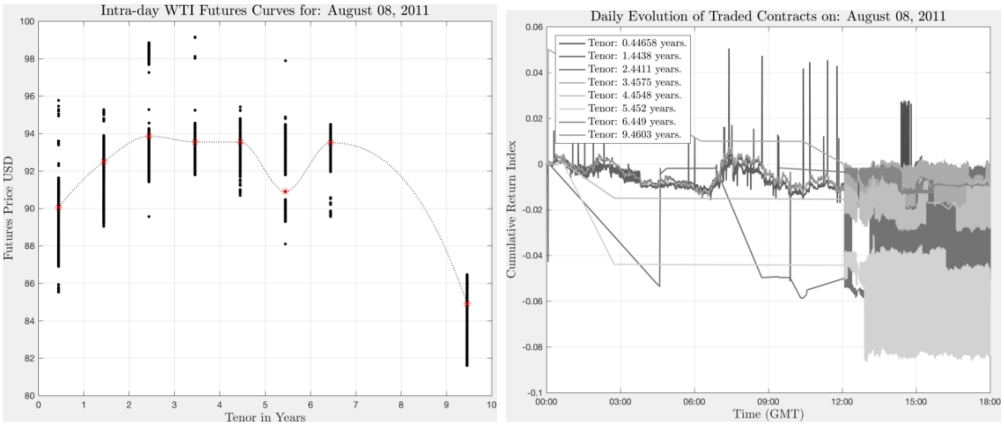


Figure 5.18: The term structure and cumulative return evolution for the WTI futures market for a single day.

5.6.1 WTI factor structure identification

The final step is to determine, the consistency of the bootstrap and, second step, is to find some indication of the underlying data generating process so we can identify the number of factors in the WTI term structure. Let $\boldsymbol{r}(t + \Delta t) = \log(\boldsymbol{P}(t + \Delta t)) - \log(\boldsymbol{P}(t))$ be the vector of returns for a given grid with time increment Δt, we always use a five minutes grid in this exercise, hence $\Delta t =$

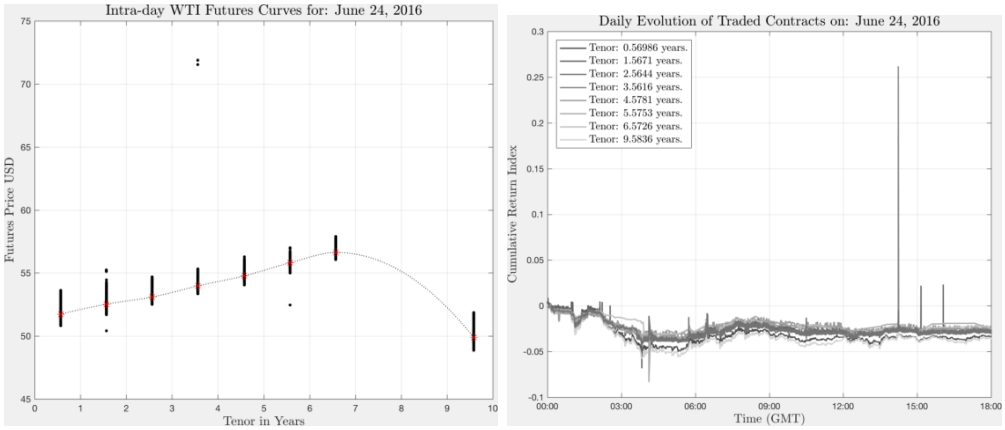


Figure 5.19: The term structure and cumulative return evolution for the WTI futures market for a single day.

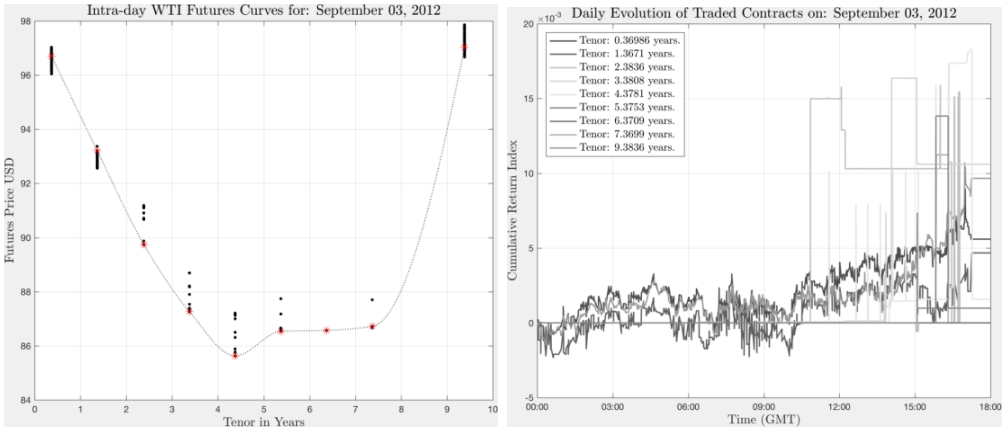


Figure 5.20: The term structure and cumulative return evolution for the WTI futures market for a single day.

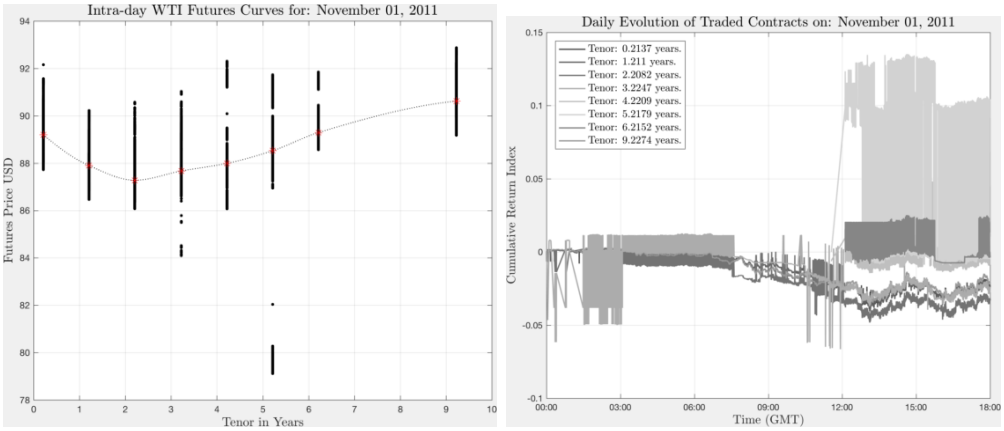


Figure 5.21: The term structure and cumulative return evolution for the WTI futures market for a single day.

$5/(24 \times 60)$ when one time increment is a single day. Let $N = (t_j - t_i)/\Delta t$, be the integer number of time increments from t_i to t_j, where t_i is the starting time and t_j is the terminal time period for each block. Where $\boldsymbol{P}(t)$ is the observed price of the vector of futures contracts at time t and is of dimension p.

Let us look at normal Chicago trading hours for the futures from 9am to 5pm weekdays and Saturdays. We ignore the limited Sunday trading and overnight and early morning. Trading days are broken up by 45 minutes of down time each day. Hence for a week we will have day 1: $(t_0, t_1, t_2, t_3)_1$, day 2: $(t_0, t_1, t_2, t_3)_2$ day 3: $(t_0, t_1, t_2, t_3)_3$, day 4: $(t_0, t_1, t_2, t_3)_4$ and day 5: $(t_0, t_1, t_2, t_3)_5$. As such the integrated covariance matrix for a five days in one week for a five minute grid would be:

$$\widehat{IV} = \sum_{d=1}^{5} IV_d, \quad \text{where, } IV_d = \sum_{t=t_0}^{t_1} \boldsymbol{r}(t+\Delta t)\boldsymbol{r}(t+\Delta t)' + \sum_{t=t_2}^{t_3} \boldsymbol{r}(t+\Delta t)\boldsymbol{r}(t+\Delta t)' \tag{5.4}$$

Let $S = IV/\sum_{d=1}^{5} N_d$ (N_d number of days) be the estimated covariance matrix,

with Σ as the 'true' data generating covariance matrix equivalent. We follow the standard identity to have:

$$\Sigma = \Lambda\Lambda' + \sigma^2 I$$

Hence, the number of bounded eigenvalues will be p minus the rank of $\Lambda\Lambda'$. Hence we can presume the following function form:

$$\log(\boldsymbol{P}(t+\Delta t)) - \log(\boldsymbol{P}(t)) = \int_t^{t+\Delta t} \Lambda \boldsymbol{f}(s) ds + \int_t^{t+\Delta t} \epsilon(s) ds \tag{5.5}$$

where $\boldsymbol{f}(s)$ is a set of uncorrelated factors of dimension k^* driven by a Brownian Semi Martingale $\mathcal{BSM}$ of unknown type and $\epsilon(s)$ is a vector of microstructure noise also drive by independent $\mathcal{BSM}$, such that:

$$\int_t^{t+\Delta t} \epsilon(s)\epsilon(s)' ds = \Delta t \sigma^2 I. \tag{5.6}$$

Whilst this second equation seems restrictive, it does force the long run microstructure noise to be uncorrelated.

Hence the eigenvalues of the integrated covariance matrix are presumed to be bounded and therefore identifiable. Of course it is impossible to directly test for this assumption as the bias correction of the eigenvalues assumes the $\mathcal{BSM}$ structure under the null. A similar problem occurs in linear regression as the OLS estimator both presumes and forces the disturbance term to be uncorrelated with the dependent variable.

An alternative approach is to attempt to directly extract the distribution of the lower block of eigenvalues under the null, for instance, the approach proposed

in Williams, Dovonon, and Taamouti [2017] directly extracts the distribution of the lower block under the null via a correlation preserving block bootstrap.

5.6.2 Results: number of factors

Using the same data documented in Chapter 2 in suction 2.4 we estimate the number of factors for the WTI futures from 2009 to 2014. This period is chosen as the number of available contracts is always sufficient to justify direct estimation of a factor model. The upperplot in Figure 5.22 presents the available number of contracts per day for these years. Notice that there are tail spikes with a very low count at regular intervals, this is the number of contracts traded on Sunday and these are ignored from the analysis.

The lower plot in Figure 5.22 presents the number of factors implied by the bootstrap (Theorem 10) test for each week in the sample. The lowest number of detected factors for the most parsimonious model criteria (in blue) is two and the highest number of factors for the most liberal selection criteria is 12. Indeed, the blue plot represents the precise stepwise test based on the power functions in the previous section (see section 5.6.1). Here we see that between two and seven factors is all that is required to summarize the data, with between three and four being more common.

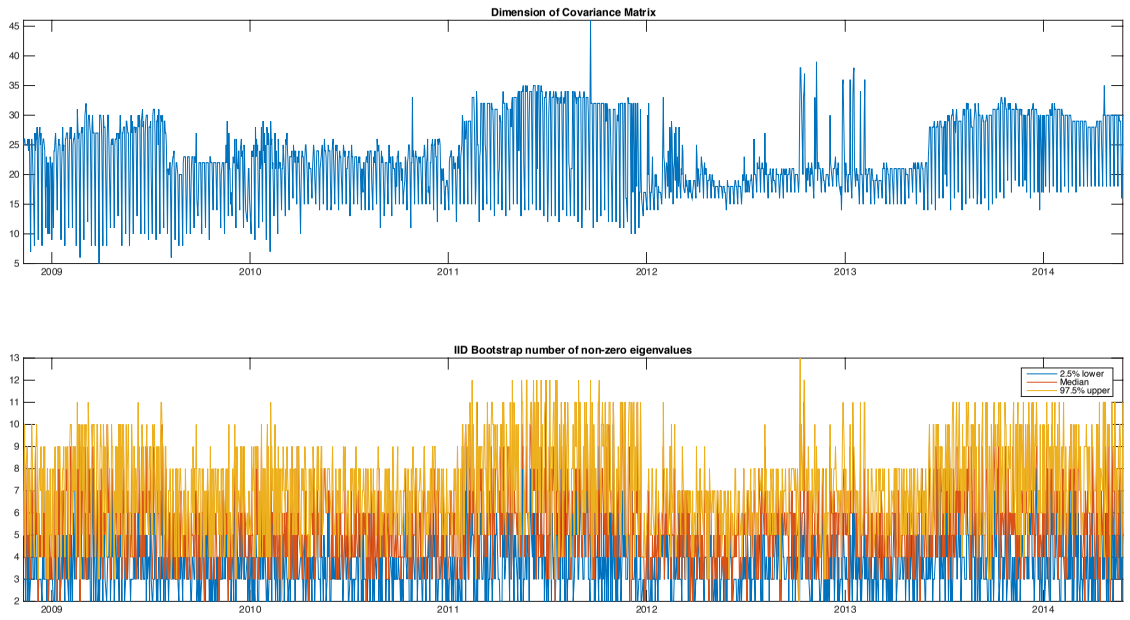


Figure 5.22: Bootstrap estimated number of factors in the WTI Term structure. The top plot represent the number of available futures contracts, distinct by tenor, available for the analysis. The bottom plot represents the number of factors detected for three quantiles from the resamples for each week over the sample period.

5.6.3 Results: quadratic variation explained

The clear objective of any futures strategy is to reduce the mark to marketet exposure for a specific delivery date. Consider the following hedging problem. At time $T*$ you have a specific position Q that needs to be physically delivered (if Q is positive then this is a forward purchase and it is negative then it is forward sale). Taking a specific position in $\mathrm{QP}_{T^*}(t)$, where $\mathrm{P}_{T^*}(t)$ is the log price of a future expiring at T^* at time t for a contract delivering at time T^* exposes the holder of the position to incremental cash flows as the prevailing spot price changes. Hence a trader will wish to take a countervailing synthetic position in a spread of other contracts to eliminate their exposure to fluctuations in $\mathrm{P}_{T^*}(t)$, this portfolio can be expressed as:

$$\Pi(t) = \mathrm{Q}(\Delta \mathrm{P}_{T^*}(t) - \boldsymbol{\beta}' \Delta \boldsymbol{P}_{-T^*}(t)) \tag{5.7}$$

where $\boldsymbol{P}_{-T^*}(t)$ is the vector of futures contracts excluding $\mathrm{P}_{T^*}(t)$. The objective of the hedge will be to minimize the variation in cash flows hence the optimal hedging ratios $\boldsymbol{\beta}$ will be

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \int_t^{T^*} (\mathrm{Q}(\Delta \mathrm{P}_{T^*}(t) - \boldsymbol{\beta}' \Delta \boldsymbol{P}_{-T^*}(t)))^2 dt \tag{5.8}$$

In a factor model basis we can simplify this to the following, let S^* be the estimated covariance matrix for all futures contracts excluding $P_{T^*}(t)$. Setting $\hat{\boldsymbol{f}}^*(t) = \hat{\boldsymbol{V}}^{*\prime} \Delta \boldsymbol{P}_{-T^*} t(t)$ where $\hat{\boldsymbol{V}}^*$ is a matrix of eigenvectors from the first $\hat{k}^*$ largest eigenvalues of S^* where $\hat{k}^*$ is the detected number of factors indicated by

our preferred test statistic.

5.6.4 Out of sample hedging results

The final analysis provides a simple comparison, over the 2011 to 2015 period is using the PCA tool to compute exposures for different hedging time frames and comparing alternative strategies. In this instance we will look at three comparisons.

1. A constant hedge, where for delivery at T^*, the hedging ratios $\boldsymbol{\beta}^* = [\beta_i^*]$ are equally distributed amongst the other available contracts, hence:

$$\beta^*_{i\in\{1,\dots,N_{C,-T^*}\}} = 1/N_{C,-T^*},$$

   where $N_{C,-T^*}$ is the count of other available contracts.

2. A constant factor model, where

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \int_t^{T^*} (\mathrm{Q}(\Delta \mathrm{P}_{T^*}(t) - \boldsymbol{\beta}' \boldsymbol{F}_{-T^*}(t)))^2 dt,$$

   where $\boldsymbol{F}_{-T^*}(t) = \boldsymbol{V}'_{3,-t} \Delta \boldsymbol{P}_{-T^*}(t)$ are the factor loadings for the first three loadings from the preceeding week.

3. A dynamic factor model, where

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \int_t^{T^*} (\mathrm{Q}(\Delta \mathrm{P}_{T^*}(t) - \boldsymbol{\beta}' \boldsymbol{F}_{-T^*}(t)))^2 dt,$$

where $\boldsymbol{F}_{-T^*}(t) = \boldsymbol{V}'_{k^*(-t)}\Delta\boldsymbol{P}_{-T^*}(t)$ are the factor loadings for the first $k^*(-t)$ loadings week on week, where $k^*(-t)$ is the indicated number of factors for the preceding week.

The R^2 of the regression of the target hedge portfolio onto the contract that needs to be hedged, usually a target maturity, such as the nearest delivery contract of some other date, such as 6 months. When the R^2 is near unity, then the hedge is essentially perfect and eliminates all variation in the target contract price. When the hedge is near zero then the hedging provides no reduction in variance. We have

Figure 5.23 presents graphically the out of sample R^2 over the available maturities for the three strategies outlined above for the 2011 to 2016 period using five minute data with weekly rebalancing for the dynamic factors. The hedging time frame runs for futures contracts from very short maturities under one year to contracts out to 10 years (of course for ten years our sample does not cover the entire lifespan of the contract).

The slope of the curve corresponds to the interpolated decrease in hedging effectiveness as maturity increases. If the curve drops to zero then the hedge does not reduce the variance of the target contracts price variation at all. The various strategies are presented in Figure 5.23 which details the R^2 with increasing maturity of the target hedge. Longer maturities are inherently more difficult to hedge (although their variance will inherently be lower due to the Samuelson effect). However, the dynamic hedging strategy using the detected number of factor rebalanced weekly provides effective hedging out to at least seven years.

The standard result that short term hedging is highly effective, with the cal-

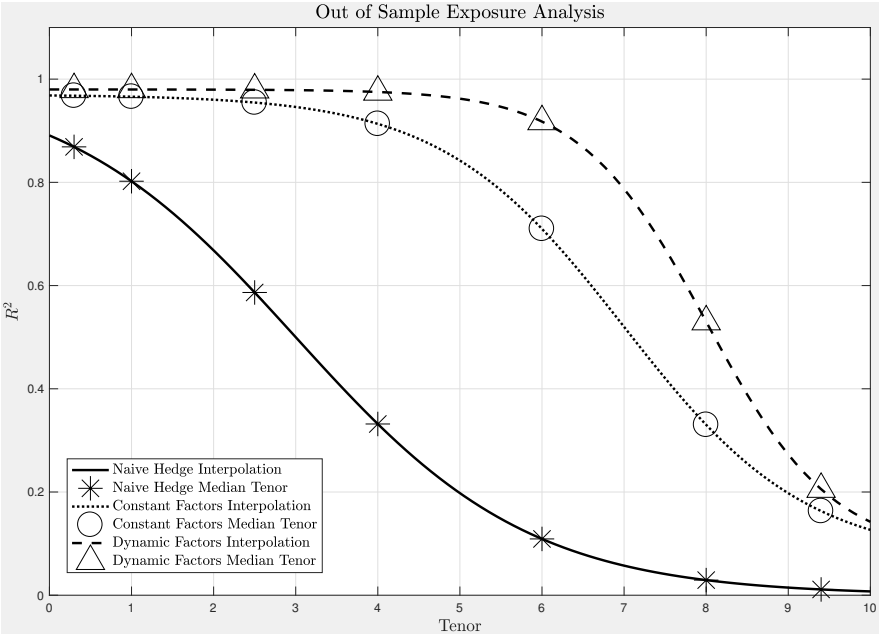


Figure 5.23: Psuedo out of sample R^2 from naive (passive), fixed factor and dynamic hedging of WTI Oil Futures.

endar spread absorbing 90% of the variation for even the Naive hedge where the objective contract is hedged by an equal weight of each of the other contracts. The dynamic portfolio allocations are conducted weekly and we have two portfolio strategies, the first using just the first three factors from the PCA analysis. This is equivalent to the exposure analysis conducted in [Aït-Sahalia and Xiu, 2018].

5.6.5 Portfolio Turnover and Economic Value

A first point to note is that every strategy within this type of hedging requires rebalancing as the maturity of the contract changes. Hence rebalancing costs are going to occur in both cases. Using the bid-ask spread to adjust for the turnover yields no difference in the value of the portfolio and this adjustment has been

reflected in the construction of Figure 5.23. As such the economic exposures illustrated in Figure 5.23 reflect, as best as possible, the ratio of variation in return explained by the hedging portfolio to the investor.

We can see that the R^2 drops toward zero for the longer tenors. This is reflective of the difficulty in fully hedging spot risk by a long (short) hold for buy (sell) side hedges. The overall speed of the change in the R^2 provides the object of interest. However, the 10 year hedge (the longest contract) still has some power when using a dynamic rebalancing approach from supervised data (i.e. adjusting the factors to the realized number inferred from the statistical test). The naive hedge drops to zero and is always lower than either of the dynamic rebalancing options.

Why does this happen? The answer can be found by returning to the results illustrated in Chapter 2 and 3. Sudden large price changes and subsequent changes in volatility precipitate from sudden changes in the order flow. We can see in the factor structure that the number of factors can shrink (over the course of minutes) at this point. Hence we have a single factor driving the majority of the variation in the entire cross section and this factor will be driven by a single contracts order-flow structure. Prices then cascade through the term structure and this is reason we see sudden declines across all maturities. In that single contract we have shown that aggressive order-flow (circa two to four standard deviations two one side of the order book) will significantly change the the mid price and hence the observed prices across the cross section of futures prices maturities.

Combining the two approaches, we can see that precipitating adjustments in overall prices reduces down to a very small number of factors and clearly a

small number of transactions and quotes and these transactions have a marked impact on the entire term structure. However, without specific identifiers on the transactions this is difficult to decompose as the limit order book data we have, herein, is anonymous.

5.7 Conclusions

we have outlined a bootstrap approach suitable for testing a very high frequency data. High-frequency data can provide many of challenges. we propose to use the VWAP-mid price to ensure that there is enough continuous variation in prices to compute the required correlations.

Our underlying model takes account of reduced rank equilibrium adjustments in price and full rank microstructure noise, the simplest case has a helpful representation in terms of signal to noise ratio.

We then used this to compute the implied rank of the covariance matrix of oil futures. The future work will be an additional work in this area will be to include the new results from the new working paper of Aït-Sahalia and Xiu [2018] who look at non-overlapping blocks, rather than the weekly sampling I have utilized in this thesis.

Chapter 6

Thesis conclusions

The use of high frequency data in asset pricing and portfolio selection is somewhat controversial. High frequency asset pricing data is noisy and there is considerable attention paid to how much information is actually contained within any signal extracted. In this thesis we have outlined a series of tools used to try and bridge the gap between the traditional approaches based around linear regression analysis and generalized method of moments and approaches founded in the non-linear and non-Gaussian stochastic processes.

The analysis is designed to look at oil futures in two dimensions. First, in Chapters 2 and 3 from the viewpoint of trade by trade and orders are submitted into the limit order book and the detection of the time that the order-book is in disequilibrium subsequent to these shocks. The results illustrate the current limit of high speed trading in determining the adjustment path of prices to new order flow. Second, in Chapters 4 and 5 we move to the cross section of futures contracts and their price evolution at high frequency.

In each case we have introduced two new statistical technologies and a new measure of order-flow imbalance for high-frequency traders across the three main empirical chapters. The technologies are designed to apply existing tools (vector autoregressions and principal component analysis) but for very high frequency data. Our object of interest is in high-frequency trading in the oil futures market, one of the most actively traded in the world.

In addition to our methodological contribution we collected a unique dataset of every best bid and ask from 1996 to 2016 and every quote order in the market from 2011 to 2016. To our knowledge, this is the first thesis to directly model the order book of a large market using a fully specified time series model specifically designed to cater for the stylized facts presented by data of this type. Our new spectral density estimator for the VAR (called the Kernal VAR) model uses a quasi spectral method of moments analysis to correct for non-normality in the underlying data set. The estimator corrects for: asynchronous recording of data, jumps and stochastic volatility. Simulation analysis suggests that the estimator is unbiased and has power given the large number of available observations, presuming that some long run relationship to the order-flow imbalance is true. The innovation in this model is that the smoothing kernel and iteration of the regression are combined to allow the model to overcome biases generated by the non-standard properties of high frequency data (local invariance and lack of time synchronization). This is then applied to high frequency data across the limit order book, allowing the direct modelling of the impact of orderflow on the price level across all levels of the book. This is the first analysis of its type.

For each of our tools we carefully assessed the ability of the tool to cope with

simulated conditions similar to those found in the market. This statistical testing provides more robust inference from the models. We then implemented the tools on actual market data from the WTI futures market traded on NYMEX.

For Chapters 2, and 3 we looked deeply (market depth 10 levels) in oil futures market for WTI. Where we applied our spectral least square estimator fitting vector autoregression, for the high frequency data. While the Monte-Carlo simulation found when the generating data process is contaminated with a complex auto and cross auto covariance, skewness and kurtosis structure. The estimator out-perform simple OLS, the OLS perfectly fails.

We checked the models simultaneous for the mid price return, and how the trader shock the order flow imbalance to generate shocks in mid price, for arbitrage positions, benefiting from the speed (around a tenth of the second) which been shown the Impulse response analysis.

The results have interesting implications for traders building algorithms to exploit adjustments in the order following shocks from new order flow. First, in Chapters 2 and 3 we show that measuring order imbalance at the milliseconds level can give near perfect guidance on the direction of the market for about 100 milliseconds, hence a trading strategy that can operate at or below 100 milliseconds will be able to generate extensive profits (as a function of the input buys and sells). This arbitrage opportunity is persistent across assets and through time and is particularly prevalent when trading in the contracts is very active. This new measure, the market depth weighted order imbalance, uses the full order book to create a series of depth measures that determine the temporal imbalance of excess supply and demand in the market. Future work, can look at determining

whether this measure works will only for actively traded contracts such as crude oil or for any type of traded asset.

In Chapters 4 and 5 we looked at term structure of WTI futures contracts and implement a bootstrap estimator to extract the latent factors structure from the co-evolution of the term structure. In Chapter 4 first we present the results on the consistency of a bootstrap estimator for the number of principal components using a likelihood ratio statistic, in opposition to the the standard information criterion or threshold approach.

The results in Chapter 5 show under hedging ratio the number of factors that been generated by the bootstrap estimated. The the power test has failed to determine the number of factors as been tested. This thesis is the first to look directly at the temporal properties of the complete limit order book and then analyses the implications for pricing over a number of timescales from HFT trading at millisecond intervals to hedging over ten years. The directions for this research are clearly endless as computing technology has now increased to a point where this type of analysis of the limit order is possible.

Appendix

Proof of Limit in Kernel VAR Model, following from on Newey West

Proof. Theorem:1

Theorem:1. Notice that

$$z_t(i,l).z_{t-s}(j,m) = \sum_{r=0}^{z}\sum_{v=0}^{z}\psi_{il}^{(r)}\psi_{jm}^{(v)}\varepsilon_{l.t-r}\varepsilon_{m.t-s-v}$$

$$(1/T)\sum_{t=1}^{T} z_t(i,l).z_{t-s}(j,m) \xrightarrow{p} E\{z_t(i,l).z_{t-s}(j,m)\}$$

$$\begin{aligned}(1/T)\sum_{t=1} T y_{it}y_{j.t-s} &\\ &= (1/T)\sum_{t=1} T\left[\mu_i + \sum_{l=1}^{n} z_t(i,l)\right]\left[\mu_j + \sum_{m=1}^{n} z_{t-s}(j,m)\right]\\ &= \mu_i\mu_j + \mu_i\sum_{m=1}^{n}\left[(1/T)\sum_{t=1}^{T} z_{t-s}(j,m)\right] + \mu_j\sum_{l=1}^{n}\left[(1/T)\sum_{t=1}^{T} z_t(i,l)\right]\end{aligned}$$

$$
\begin{aligned}
&+\sum_{l=1}^{n}\sum_{m=1}^{n}\left[(1/T)\sum_{t=1}^{T}z_t(i,l)z_{t-s}(j,m)\right]\\
&\xrightarrow{p}\mu_i\mu_j+\mu_i\sum_{m=1}^{n}E[z_{t-s}(j,m)]+\mu_j\sum_{l=1}^{n}E[z_t(i,l)]\\
&+\sum_{l=1}^{n}\sum_{m=1}^{n}E[z_t(i,l)z_{t-s}(j,m)]\\
&=E\left\{\left[\mu_i+\sum_{l=1}^{n}z_t(i,l)\right]\left[\mu_j+\sum_{m=1}^{n}z_{t-s}(j,m)\right]\right\}\\
&=E[y_{it}y_{j.t-s}]
\end{aligned}
$$

■

Functions and codes chapters 2 and 3

Ansynchronous VAR model for chapter 2 and chapter 3

The following matlab codes are designed to run the spectral VAR estimator from this chapter. The function `AsynchronousRegression.m` is the main file and implements the spectral estimator. `Autocovariance.m` and `CrossCovariance.m` compute the kernel auto and cross covariances for the multivariate estimator. `impulseResponseFunction.m` plots the resulting impulse response function with `ImpulseErrorBounds.m`, `CompanionMatrix.m` and `createMatchedLagMatrix.m` as utilities. `OrderBookPlot.m` is a utility that plots the order book. All codes are by the author.

```
function output=AsynchronousRegression(Y,info)


%T is a cell array of vectors such that T{1,i} = T1 x 1 vector
%Y is a cell array of data already transformed and de-meaned such
    that X{1,i} =
%T1 x 1 vector

%info is a structure containing the various information needed to
%run the analysi
%info.lags is the number of lags in the VAR model
%info.H is the nuisance parameter for the realized kernels
%info.saveData is a 0 or 1 flag, 1 saves the synchronized data
    matrices and
```

```
13 %tick times 0 stops this from being recorded in the output
14 %info.coption =1 sets the constant to true
15
16 %output is a data structure
17 %if info.saveData == 1
18 %output.t is the T+nlag x 1 vector of synchronized tick times
19 %output.X is the T x N*nlags matrix of lagged explanatory variables.
20 %output.Y is the T x N*nlags matrix of contemporaneous dependent
       variables.
21 %
22 %
23 %output."kernel".Pi is the matrix of coeeficients (N x N*nlags)
24 %output."kernel".CovPi is the covariance matrix of coefficients (N*
       nlags x N*nlags)
25 %output."kernel".sePi is the matrix of standard erros (N x N*nlags)
26 %if info.saveData == 1
27 %output."kernel".U is the T x N matrix of VAR residuals
28 %output."kernel".roots is the N*nlag list of roots of the VAR
29 %
30 %REMEMBER Matlab will not allow you to concatenate arrays with
       different
31 %field orderings, so either run the wrapper loop with info.saveData
       == 1 or
32 %info.saveData == 0 do not try to mix both or you will get a
       concatenation
33 %error
34 %
35 output=info;
36 N=size(Y,2);
37 nlag=info.lags;
38 cinf=info.coption;
39 [X,Y]=createMatchedLagMatrix(Y,nlag);
40 %add a constant
41 if cinf==1
42     X=[X ones(length(X),1)];
43 end
44 TT=length(Y);
45 output.T=TT;
46 %important this overwrites the input X and is used from now on.
47 kernels={'parzen';'qspec';'fejer';'tukey';'bhnls'};
48 %run the regression for each kernel;
49
50 %use the saveData flag to optionally collect the matched data
51 %for very large microstructure data this should be set to 0
52 %unless absolutely necessary for other robustness checks
53 if info.saveData==1
54     output.X = X;%save the lagged explanatory variables
55     output.Y = Y;%save the dependent variables
56 end
```

```
57
58 H=info.H;
59 if H == 0 %this is the flag to stop the kernel being used
60     %whilst it is a bit wasteful to repeat this five times, it is
           fast even
61     %if the data is quite big
62     info.kernel='parzen';
63     XX=X'*X;
64     XY=X'*Y;
65     %compute the coefficients
66     Pi = inv(XX)*XY;%#ok<MINV>
67     %collect the residuals
68     U = Y - X*Pi;
69     %call the realized kernel again to compute the standard errors.
70     %Sigma=MultivariateRealizedKernel(U,info);
71     Sigma=cov(U);
72     %compute the coefficients covariance matrix
73     covPi = kron(Sigma,inv(XX./TT));
74     %extract the standard errors, reshape
75     sePi = reshape(sqrt(diag(covPi)),size(Pi))./sqrt(TT);
76     c=zeros(N,1);
77     if cinf==1
78         c=Pi(end,:)';
79         Pi(end,:)=[];
80     end
81     if nlag==1
82         roots=eig(Pi);
83     else
84         Ntop = N*nlag;
85         F = [Pi' zeros(N,N);eye(Ntop) zeros(Ntop,N)];
86         roots=eig(F);
87     end
88     %save the data to the output structure
89     output.parzen.Pi=Pi;
90     output.parzen.covPi=covPi;
91     output.parzen.sePi=sePi;
92     output.parzen.constant=c;
93     if info.saveData==1
94         output.parzen.U=U;
95     end
96     output.parzen.roots=roots;
97     output.parzen.Sigma=Sigma;
98
99
100    output.qspec.Pi=Pi;
101    output.qspec.covPi=covPi;
102    output.qspec.sePi=sePi;
103    output.qspec.constant=c;
104    if info.saveData==1
```

```
105             output.qspec.U=U;
106         end
107         output.qspec.roots=roots;
108         output.qspec.Sigma=Sigma;
109 
110         output.fejer.Pi=Pi;
111         output.fejer.covPi=covPi;
112         output.fejer.sePi=sePi;
113         output.fejer.constant=c;
114         if info.saveData==1
115             output.fejer.U=U;
116         end
117         output.fejer.roots=roots;
118         output.fejer.Sigma=Sigma;
119 
120         output.tukey.Pi=Pi;
121         output.tukey.covPi=covPi;
122         output.tukey.sePi=sePi;
123         output.tukey.constant=c;
124         if info.saveData==1
125             output.tukey.U=U;
126         end
127         output.tukey.roots=roots;
128         output.tukey.Sigma=Sigma;
129 
130 
131         output.bhnls.Pi=Pi;
132         output.bhnls.covPi=covPi;
133         output.bhnls.sePi=sePi;
134         output.bhnls.constant=c;
135         if info.saveData==1
136             output.bhnls.U=U;
137         end
138         output.bhnls.roots=roots;
139         output.bhnls.Sigma=Sigma;
140 
141     elseif H>0
142         for i=1:length(kernels)
143             info.kernel=char(kernels{i,1});
144             XX=MultivariateRealizedKernel(X,info);
145             XY=MultivariateRealizedCrossVariation(X,Y,info);
146             %compute the coefficients
147             Pi = inv(XX)*XY; %#ok<MINV>
148             %collect the residuals
149             U = Y - X*Pi;
150             %call the realized kernel again to compute the standard
                    errors.
151             Sigma=MultivariateRealizedKernel(U,info);
152             %compute the coefficients covariance matrix
```

```
153         covPi = kron(Sigma,inv(XX./TT));
154         %extract the standard errors, reshape
155         sePi = reshape(sqrt(diag(covPi)),size(Pi))./sqrt(TT); %#ok<
                NASGU>
156         c=zeros(N,1);
157         if cinf==1
158             c=Pi(end,:)';
159             Pi(end,:)=[];
160         end
161         if nlag==1
162             roots=eig(Pi); %#ok<NASGU>
163         else
164             %compute the companion matrix
165             Ntop = N*nlag;
166             F = [Pi' zeros(N,N);eye(Ntop) zeros(Ntop,N)];
167             roots=eig(F); %#ok<NASGU>
168         end
169         %collect the outputs by commiting the eval that men do ...
170         str=['output.',char(kernels{i,1}),'.Pi=Pi;'];eval(str);
171         str=['output.',char(kernels{i,1}),'.covPi=covPi;'];eval(str)
                ;
172         str=['output.',char(kernels{i,1}),'.sePi=sePi;'];eval(str);
173         if info.saveData==1
174             str=['output.',char(kernels{i,1}),'.U=U;'];eval(str);
175         end
176         str=['output.',char(kernels{i,1}),'.roots=roots;'];eval(str)
                ;
177         str=['output.',char(kernels{i,1}),'.Sigma=Sigma;'];eval(str)
                ;
178         str=['output.',char(kernels{i,1}),'.constant=c;'];eval(str);
179     end
180 end
```

```
1  function A=Autocovariance(X,j)
2
3  %jth order auto covariance matrix
4  %of the T x n data matrix X
5  %YOU SHOULD DEMEAN THE MATRIX X PRIOR TO RUNNING THIS FUNCTION
6
7  A = (X(1+abs(j):end,:)'*X(1:end-abs(j),:));
8  if sign(j)==-1
9      A = A';
10 end
11
12
13
14
15
16
```

```
17
18
19 % dims=size(X);
20 % if j==0
21 %     A=(X'*X)./(dims(1)+1);
22 % elseif j<0
23 %     A=zeros(dims(2),dims(2),dims(1)-j);
24 %     for i=1:dims(1)+j
25 %         xi=X(i,:)';
26 %         xj=X(i-j,:)';
27 %         A(:,:,i)=xi*xj';
28 %     end
29 %     A=squeeze(sum(A,3))./(dims(1)-abs(j)-1);
30 % elseif j>0
31 %     A=zeros(dims(2),dims(2),dims(1)-j);
32 %     for i=j+1:dims(1)
33 %         xi=X(i,:)';
34 %         xj=X(i-j,:)';
35 %         A(:,:,i)=xi*xj';
36 %     end
37 %     A=squeeze(sum(A,3))./(dims(1)-abs(j)-1);
38 % end
39 %
```

```
1  function A=CrossCovariance(X,Y,j)
2
3  %the auto-cross-covariation
4  %X is a T x n matrix
5  %Y is a T x m matrix
6  %j is an integer
7  %A is a n x m cross
8  %YOU SHOULD DEMEAN the mnatrices X and Y before running this
       function
9
10 %This procedure is for computational speed-up to check
11 %consistency do the following:
12 %
13 % XY = [X Y]
14 %run RK = MultivariateRealizedKernel(XY,info) %info should be the
       same as in
15 %this function
16 %
17 %The regression is inv(RK(1:n,1:n))*RK(n+1:2*n,n+1:n+m)
18
19 %this is a cross covariance matrix so there is no need to transpose
       it for
20 %negative auto-covariances as these are from the bottom partition:
21
22 %
```

```
% S = [C A]
%     [A'D]
%
%where S = autocov([X Y],j)

A = (X(1+abs(j):end,:)'*Y(1:end-abs(j),:));



% [T,n]=size(X);
% [~,m]=size(Y);

% if j==0
%     A=(X'*Y)./(T+1);
% elseif j<0
%     A=zeros(n,m,T-j);
%     for i=1:T+j
%         xi=X(i,:)';
%         xj=Y(i-j,:)';
%         A(:,:,i)=xi*xj';
%     end
%     A=squeeze(sum(A,3))./(T-abs(j)-1);
% elseif j>0
%     A=zeros(n,m,T-j);
%     for i=j+1:T
%         xi=X(i,:)';
%         xj=Y(i-j,:)';
%         A(:,:,i)=xi*xj';
%     end
%     A=squeeze(sum(A,3))./(T-abs(j)-1);
% end
```

```
function [IRF]=impulseResponseFunction(Pi,s,flag)

%Fast Impulse Response Generator can be called repeatedly for simulated
%S.E.s
%for Realized VAR
%Pi is nr x n the matrix of autoregressive coefficients (without the intercept)
%s is the number of steps
%Sigma is the covariance of the errors (identity eye(n) for unit shocks)
%Pi should already have the intercept row stripped out

%IRF is the impulse responses in an n x n x s array;
%Send Sampled Pis and sort for error bounds
%Flag ==1 Cholesky
```

```
%Flag ==2 Generalized


if nargin==3 && flag==1
    %delta simulation method (can be faster than matrix squares).
    n = size(Pi,2);
    m = size(Pi,1);
    r = m/n;
    IRF=zeros(n,n,s);
    for i=1:n
        d = zeros(n,1);
        d(i)= 1;
        for j=1:s
            if j==1
                IRF(i,:,j)=d;
            else
                if j<=r
                    jl=j-1;
                    LIRF=reshape(squeeze(IRF(i,:,1:jl)),[],1);
                    PiL = Pi(1:(n*jl),:);
                    IRF(i,:,j)=PiL'*LIRF;
                else
                    jl=j-1;
                    LIRF=reshape(squeeze(IRF(i,:,j-r:jl)),[],1);
                    IRF(i,:,j)=Pi'*LIRF;
                end
            end
        end
    end
elseif nargin==2 || flag>1
    %Generalised IRFs
    n=size(Pi,2);
    m=size(Pi,1);
    r = m/n;
    if r==1
        F=Pi;
    else
        F=[Pi' zeros(n,n);eye(n*r) zeros(n*r,n)];
    end
    %cycle through the G-IRFs
    IRF=zeros(n,n,s);
    for i=1:s
        Fi=F^i;
        Fii=Fi(1:n,1:n);
        IRF(:,:,i)=Fii;
    end
end
```

```
function [F]=CompanionMatrix(Pi,n,nlag)


F1 = Pi';F2 = eye(n*(nlag-1));F3 = zeros(n*(nlag-1),n);
F = [F1;F2 F3];
```

```
function [Y,stats]=makeOrderBookData(Bid,Ask,spec)

%This function makes orderBook data to spec

%spec.normalize = 0 or 1, 0 leaves the data 1, demeans and divides
    by the
%standard deviation

% spec =
%
%                    normalize: 1 or 0
%                  stalePrices: 1 or 0
%              returnCalcLevel: 1 to 5 (or max levels)
%     numberBidAskSpreadLevels: 1 to 5 (or max levels)
%           numberVolumeRatios: 1 to 5 (or max levels)

nflag = spec.normalize;
rlevel = spec.returnCalcLevel;
slevel = spec.numberBidAskSpreadLevels;
vlevel = spec.numberVolumeRatios;
stalePrices = spec.stalePrices;



NL = max([ rlevel slevel vlevel]);

PA = Ask.P(:,1:NL);
PB = Bid.P(:,1:NL);
VA = Ask.V(:,1:NL);
VB = Bid.V(:,1:NL);

%quick sanity check on the data
T=length(PA);
nind=sum(isnan(PA))./T;
zind=sum(PA==0)./T;
ntol=find(nind>1/3);
ztol=find(zind>1/3);
minNL = min([ntol ztol])-1;

stats.dataIssueflag=0;
if minNL<NL
    if minNL>=1
        NL = min([NL minNL]);
```

```
43          slevel = min([NL slevel]);
44          vlevel = min([NL vlevel]);
45          disp('Insufficient data in selected  levels > 1, selecting a
                lower number');
46          stats.dataIssueflag=1;
47      else
48          error('Data does not have enough observations at L1 to be
                viable');
49      end
50  end
51
52
53  dvA=Ask.dv;
54  dvB=Bid.dv;
55
56  %check the timestamps as this causes
57  %huge issues with the
58  iA = (1:length(dvA))';
59  [uA,ii]=unique(dvA);
60  iiA=iA(ii);
61  dvA=interp1(iiA,uA,iA);
62  iii=find(isnan(dvA)+[0;(diff(dvA)==0)]);
63  dvA(iii)=[];
64  PA(iii,:)=[];
65  VA(iii,:)=[];
66  %NA(iii,:)=[];
67
68  iB = (1:length(dvB))';
69  [uB,ii]=unique(dvB);
70  iiB=iB(ii);
71  dvB=interp1(iiB,uB,iB);
72  iii=find(isnan(dvB)+[0;(diff(dvB)==0)]);
73  dvB(iii)=[];
74  PB(iii,:)=[];
75  VB(iii,:)=[];
76  %NB(iii,:)=[];
77
78  dv=dvB;%set the bids to be the master date vector
79  for i=1:NL
80      [PA(:,i)]=QuickcleanHF_Data(PA(:,i),dvA);
81      [PB(:,i)]=QuickcleanHF_Data(PB(:,i),dvB);
82      [VA(:,i)]=QuickcleanHF_Data(VA(:,i),dvA);
83      [VB(:,i)]=QuickcleanHF_Data(VB(:,i),dvB);
84      %[NA(:,i)]=QuickcleanHF_Data(NA(:,i),dvA);
85      %[NB(:,i)]=QuickcleanHF_Data(NB(:,i),dvB);
86
87      %rebase the asks to the contemparaneous bids
88      PA(:,i)=interp1(dvA,PA(:,i),dv);
89      VA(:,i)=interp1(dvA,VA(:,i),dv);
```

```
90      % NA(:,i)=interp1(dvA,NA(:,i),dv);
91      %check the data for gaps again
92      [PA(:,i)]=QuickcleanHF_Data(PA(:,i),dv);
93      [VA(:,i)]=QuickcleanHF_Data(VA(:,i),dv);
94      %[NA(:,i)]=QuickcleanHF_Data(NA(:,i),dv);
95  end
96
97  MP = (nansum(PA.*VA,2)./nansum(VA,2) + nansum(PB.*VB,2)./nansum(VB
        ,2))./2;
98  ret = [0;diff(log(MP))];
99
100 Y = zeros(length(ret),1 + slevel + vlevel);
101 Y(:,1) = ret;
102 for i=1:slevel
103     Y(:,i+1) = log(PA(:,i)) - log(PB(:,i));
104 end
105
106 for i=1:vlevel
107     Y(:,i+1+slevel) = log(VA(:,i)) - log(VB(:,i));
108 end
109
110 if stalePrices
111     ind = find(ret==0);%find the null returns
112 end
113 Y(ind,:)=[];
114 if nflag
115     m=nanmean(Y);
116     s=nanstd(Y);
117     M = repmat(m,length(Y),1);
118     S = repmat(s,length(Y),1);
119     stats.mean = m;
120     stats.std = s;
121     Y = (Y - M)./S;
122 else
123     stats.mean = zeros(1,size(Y,2));
124     stats.std = ones(1,size(Y,2));
125 end
126
127 stats.normalize = nflag;
128 stats.returnCalcLevel = rlevel;
129 stats.numberBidAskSpreadLevels = slevel;
130 stats.numberVolumeRatios = vlevel;
131 stats.stalePrices = stalePrices;
```

```
1  function [X,Y]=createMatchedLagMatrix(Y,nlag)
2
3  %Y is a T x n block of data assumes that Y(end,:) is the latest
4  %observations and Y(:,1) is the oldest observations.
5  %nlag is a number of desired lags
```

```
%X is the explanatory variables
%Y is the explanatory


[T,n]=size(Y);
X = zeros(T-nlag,n*nlag);
ii = [1:n:n*nlag;n:n:n*nlag]';
for i=1:nlag
    X(:,ii(i,1):ii(i,2))=Y(nlag+1-i:T-i,:);
end
Y = Y(nlag+1:T,:);
```

```
function OrderBookPlot(Ask,Bid)

%This is a plotter

%
% Ask/Bid array should be =
%
%          dv: Date vector of asks
%           P: Prices x Level data set
%           V: Volume x Level data set
%           N: Number x Level data set
%         day: Date
%         RIC: RIC from database
%        type: {'Market Depth'}
%      offset: GMT offset in hours
%      header: Header from database (for error checking)
```

```
PA=Ask.P;
PB=Bid.P;
VA=Ask.V;
VB=Bid.V;
NA=single(Ask.N);
NB=single(Bid.N);
dvA=Ask.dv;
dvB=Bid.dv;
VA(find(isnan(VA)))=0;
VB(find(isnan(VB)))=0;
VA=cumsum(VA,2);
VB=cumsum(VB,2);
NA=cumsum(NA,2);
NB=cumsum(NB,2);
indA=find(PA==0);
indB=find(PB==0);
```

```
PA(indA)=NaN;
PB(indB)=NaN;
indA=find(isnan(PA));
indB=find(isnan(PB));
VA(indA)=NaN;
VB(indB)=NaN;
NA(indA)=NaN;
NB(indB)=NaN;
[C]=unique(dvA);

NumPA = sum(~isnan(PA));


colmat=[0  0  1
        1  0  0
        0  1  0
        1  0  1
        0  0  0
        0  1  1];

set(groot,'defaultAxesColorOrder',colmat);

scrsz=get(0,'ScreenSize');
warning off all
figure('position',scrsz);
orient landscape;
disp('Figure Created Press Any Key to Continue');
pause(2);
subplot(3,1,1);
h=plot(dvA,PA);hold on;
ax = gca;
ax.ColorOrderIndex = 1;
plot(dvB,PB);hold on;grid on;datetick('x');hold off
set(gca,'fontsize',6);
g=legend(h,{['Level 1: ',num2str(NumPA(1))];...
     ['Level 2: ',num2str(NumPA(2))];...
     ['Level 3: ',num2str(NumPA(3))];...
     ['Level 4: ',num2str(NumPA(4))];...
     ['Level 5: ',num2str(NumPA(5))]});
set(g,'location','best','fontsize',6);
tname=['\bf{',char(Ask.RIC),' Order Book Ask and Bid Prices For ',
    datestr(unique(floor(dvA))),'}'];
H=title(tname);set(H,'fontsize',14);

subplot(3,1,2);
h=plot(dvA,VA);
ax = gca;
ax.ColorOrderIndex = 1;
hold on;plot(dvB,-VB);hold on;datetick('x');grid on;hold off
```

```
%legend(h,{'Level 1';'Level 2';'Level 3';'Level 4';'Level 5'});
set(gca,'fontsize',6);
tname=['\bf{',char(Ask.RIC),' Ask and Bid Order Book Volume For ',
    datestr(unique(floor(dvA))),'}'];
H=title(tname);set(H,'fontsize',14);


subplot(3,1,3);
h=plot(dvA,NA);hold on;
ax = gca;
ax.ColorOrderIndex = 1;
plot(dvB,-NB);hold on;datetick('x');grid on;hold off
%legend(h,{'Level 1';'Level 2';'Level 3';'Level 4';'Level 5'});
set(gca,'fontsize',6);
tname=['\bf{',char(Ask.RIC),' Order Book Sellers and Buyers For ',
    datestr(unique(floor(dvA))),'}'];
H=title(tname);set(H,'fontsize',14);
xin=get(gca,'xtick');



disp('Press Any Key to Commit to Disk');
pause(2);
fname=['OrderBookPlot_',char(Ask.RIC),'_',datestr(now,30),'.eps'];
saveas(gcf,fname,'psc2');
close all;
warning on all
```

```
function [IRF,IRFl,IRFu] = ImpulseErrorBounds(Pi,Sigma,s,NB,T,ca)
%UNTITLED Summary of this function goes here
%   Detailed explanation goes here
%Monte-carlo simulation of the error bounds of the IRF for Pi
%Pi is a matrix of VAR lag coefficients.
%Sigma is the covariance matrix
%s is the number of steps in the IRF
%NB is the number of replications
%T is the sample size for the simulation
%ca is the desired confidence bounds default is 95%

if nargin==5
    ca=0.025;
end
n=size(Pi,2);
m=size(Pi,1);
nlag=m/n;
Fii=WaldRepresentationVAR(Pi,nlag+1);
Q=chol(Sigma);
Pical=zeros(n,n,nlag);
ns=1:n:m;
```

```
ne=n:n:m;
for i=1:nlag
    Pical(:,:,i)=Pi(ns(i):ne(i),:);
end

disp('Starting Bootstrap Loop')
for k=1:NB
    E = randn(2*T,n);
    U = E*Q;
    %generate the initial data from the VMA sequence
    Y=U;
    for i=2:2*T
        if i<=nlag+1
            yi=U(i,:)';
            for j=1:i-1
                Psi=squeeze(Fii(1:n,1:n,j));
                yi=yi+Psi*U(i-j,:)';
            end
        else
            yi=U(i,:)';
            for j=1:nlag
                Pii=squeeze(Pical(:,:,j));
                yl=Y(i-j,:)';
                yi=yi+Pii*yl;
            end
        end
        Y(i,:)=yi';
    end
    Y=Y(T+1:end,:);
    [X,Y]=createMatchedLagMatrix(Y,nlag);
    Piik = X\Y;
    Fiik=WaldRepresentationVAR(Piik,s);
    outputMC(k).Pi=Piik;
    outputMC(k).IRF=Fiik;
    disp(['Completed: ',num2str(k)]);
end
[IRF,IRFl,IRFu]=sortIRFBounds(outputMC,ca);
end
```

.1 Functions and Codes Chapter 4

.1.1 BootStrapCorrection

This is the main function that runs the bootstrap proposed in this chapter. The inputs is Y a block of regularly spaced data, where $n^{-1}Y'Y$ is full rank and "`nboot`" is the number of

bootstrap replications.

```
function [L,BL,critLi,critBLi]=BootStrapCorrection(Y,nboot)

% L is the test statistic
% BL is the short sample adjusted statistic
% critL is the bootstrap critical value for L
% critBL is the bootstrap critical value for BL

[N,m] = size(Y);
Ybar = repmat(mean(Y),N,1);
n = N - 1;
S = Y'*Y./n;
[L,BL,~]=TestStatistic(Y);%generate the classical statistic and
    bounds
[V,l]=rootSorter(S);
% generate the m-2 condidate null matrices
Qdagger = zeros(m-1,m,m);
for k=1:m-1
    q = m - k;
    lbar = sum(l(k+1:end))/q;
    vbar = sum(V(:,k+1:end),2)/q;%by column
    ldagger = l;%(k+1:end)
    ldagger(k+1:end) = lbar;
    Vdagger = V;
    Vdagger(:,k+1:end) = repmat(vbar,1,q);
    Sdagger = V*diag(ldagger)*inv(V);
    Qdagger(k,:,:) = real(chol(Sdagger));
end

% simulate under the null.
Edraw = randn(nboot,N,m);
critLi = zeros(nboot,m);
critBLi = zeros(nboot,m);
parfor j=1:nboot
    %tic
    E = squeeze(Edraw(j,:,:));%use the same random numbers for each
        null
    for k=1:m-1
        Qd = squeeze(Qdagger(k,:,:));
        Ystar = Ybar + E*Qd;
        %[Lstar,BLstar,~]=TestStatistic(Ystar);
        Sstar = cov(Ystar);
        l = sort(eig(Sstar),'descend');
        q = m-k;
        lbar = mean(l(k+1:m));
%         for i=k+1:m
%             lbar = real(lbar + l(i)./q);
%         end
%         Vk = 1;
```

```
%           for i=k+1:m
%               Vk = real(Vk.*l(i));
%           end
          Vk = prod(l(k+1:m));
          Vk = Vk./(lbar.^q);
          Lstar = -n.*log(Vk);
          BLstar = -(n - k - (2*q^2+q+2)./(6*q)).*log(Vk);
          critLi(j,k) = Lstar;
          critBLi(j,k) = BLstar;
      end
      %toc
end
critL = prctile(critLi,95)';
critBL = prctile(critBLi,95)';
```

```
function [H,FH]=powerFunctionAnalysis(N,m,k,NREP,sigma,nboot)

rej = zeros(NREP,m);
frej = zeros(NREP,m);
Brej = zeros(NREP,m);
Bfrej = zeros(NREP,m);

bootrej = zeros(NREP,m);
bootfrej = zeros(NREP,m);
bootBrej = zeros(NREP,m);
bootBfrej = zeros(NREP,m);

parfor i=1:NREP
    [Y,~]=dgp_function(N,m,k,sigma);
    [L,BL,CritVal]=TestStatistic(Y);
    [BootL,BootBL,~]=iidBootStrap(Y,nboot);
    Brej(i,:) = double(BL<CritVal);
    Bfrej(i,:) = double(BL>=CritVal);
    rej(i,:) = double(L<CritVal);
    frej(i,:) = double(L>=CritVal);
    bootBrej(i,:) = double(BootBL<CritVal);
    bootBfrej(i,:) = double(BootBL>=CritVal);
    bootrej(i,:) = double(BootL<CritVal);
    bootfrej(i,:) = double(BootL>=CritVal);
end
scrz = get(0,'screensize');
FH1 = figure('position',ceil(scrz/2),'color','w');
[Yrej,Yfrej,YBrej,YBfrej,YrejBoot,YfrejBoot,YBrejBoot,YBfrejBoot]=
    powerCalc(rej,frej,Brej,Bfrej,bootrej,bootfrej,bootBrej,bootBfrej
    );
H1=plotPowerFunction(m,k,Yrej,Yfrej,YBrej,YBfrej);
G = title('Asymptotic and Short Sample Adjustment');
set(G,'interpreter','latex','fontsize',16);

```

```
scrz = get(0,'screensize');
FH2 = figure('position',ceil(scrz/2),'color','w');
H2=plotPowerFunction(m,k,YrejBoot,YfrejBoot,YBrejBoot,YBfrejBoot);
G = title('Bootstrap with and without sample adjustment');
set(G,'interpreter','latex','fontsize',16);
H = [H1;H2];
FH = [FH1;FH2];
```

```
function [L,BL,CritVal]=TestStatistic(Y)

[N,m] = size(Y);
n = N-1;
S = Y'*Y./n;%sample covariance
l = sort(eig(S),'descend');
L = zeros(m,1);
BL = zeros(m,1);
CritVal = L;
for k=1:m
    q = m-k;
    lbar = 0;
    for i=k+1:m
        lbar = real(lbar + l(i)./q);
    end
    Vk = 1;
    for i=k+1:m
        Vk = real(Vk.*l(i));
    end
    Vk = Vk./(lbar.^q);
    L(k) = -n.*log(Vk);
    BL(k) = -(n - k - (2*q^2+q+2)./(6*q)).*log(Vk);
    dgf = (q+2).*(q-1)./2;
    CritVal(k) = chi2inv(0.95,dgf);
end
```

```
function eigenvalueDistributionImages

%
% bias corrected bootstrao
%

[N,m] = size(Y);
if N>m
    n = N - 1;
else
    error('This model will only work with N > m')
end
mu = repmat(mean(Y),N,1);
S = cov(Y);
lsample = sort(eig(S),'descend');
```

```
E = randn(nboot,N,m);
Q = chol(S);
%generate artificial data under the null.
err = zeros(nboot,m);
for i=1:nboot
    Estar = squeeze(E(i,:,:));
    Ystar = Estar*Q + mu;
    lstar = sort(eig(cov(Ystar)),'descend');
    err(i,:) = lstar-lsample;
end
figure('color','w')
plot(sort(err(:,[30 40 50])),(1:nboot)'./nboot,'linewidth',2);
G = legend({'Latent Root No. 30','Latent Root No. 40','Latent Root
    No. 50'});
set(G,'interpreter','latex','fontsize',12,'location','best');
grid on
G = title('Simulated Error Distribution of Eigenvalues');
set(G,'interpreter','latex','fontsize',16);
G = xlabel('$l_i - \lambda_i$');
set(G,'interpreter','latex','fontsize',14);
G = ylabel('EDF')
set(G,'interpreter','latex','fontsize',14);
export_fig error_example.pdf

% correction = median(err)';
% l = lsample+correction;
% %now recompute the test statistic
% BootL = zeros(m,1);
% BootBL = zeros(m,1);
% BootCritVal = BootL;
% for k=1:m
%     q = m-k;
%     lbar = 0;
%     for i=k+1:m
%         lbar = lbar + l(i)./q;
%     end
%     Vk = 1;
%     for i=k+1:m
%         Vk = Vk.*l(i);
%     end
%     Vk = Vk./(lbar.^q);
%     BootL(k) = -n.*log(Vk);
%     BootBL(k) = -(n - k - (2*q^2+q+2)./(6*q)).*log(Vk);
%     dgf = (q+2).*(q-1)./2;
%     BootCritVal(k) = chi2inv(0.95,dgf);
% end
%
```

Bibliography

Admati, A. R. and P. Pfleiderer (1988). A theory of intraday patterns: Volume and price variability. *The Review of Financial Studies 1*(1), 3–40. 113

Admati, A. R. and P. Pfleiderer (1989). A theory of intraday and day-of-the-week mean effects. *The Review of Financial Studies 2*(2), 189–223. 113

Aït-Sahalia, Y. and D. Xiu (2015). Principal component analysis of high frequency data. Technical report, Technical report, Princeton University. 28

Aït-Sahalia, Y. and D. Xiu (2018). Principal component analysis of high-frequency data. *Journal of the American Statistical Association*, 1–17. 13, 97, 99, 105, 109, 110, 113, 114, 115, 130, 160, 161, 162, 189, 192

Alquist, R. and O. Gervais (2013). The role of financial speculation in driving the price of crude oil. *The Energy Journal 34*(3), 35. 26

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001). The distribution of realized exchange rate volatility. *Journal of the American statistical association 96*(453), 42–55. 40

Anderson, R. (1983). *The Industrial Organization of Futures Markets*. Columbia University. Center for the Study of Futures Markets. 31, 37, 38, 39

Anderson, T. W. (1959). On asymptotic distributions of estimates of parameters of stochastic difference equations. *The Annals of Mathematical Statistics*, 676–687. 100, 107, 118, 121, 161, 172

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics 34*(1), 122–148. 104, 105, 107, 117, 118, 139, 141, 142, 161

Apostol, T. M. (1969). *Calculus: Mullti Variable Calculus and Linear Algebra, with Applications to Diffferential Equations and Probability*. John Wiley & Sons. 119, 147

Audrino, F. and F. Corsi (2008). Realized covariance tick-by-tick in presence of rounded time stamps and general microstructure effects. Working Paper Series 4, University of St. Gallen. 114

Balakrishnan, N. (2006). *Continuous multivariate distributions*. Wiley Online Library. 148

Barndorff-Nielsen, O. and N. Shephard (2004a). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica 72*(3), 885–925. 94

Barndorff-Nielsen, O. E. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(2), 253–280. 40

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica 76*(6), 1481–1536. 41, 58, 64, 65, 66

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2009a). Realized kernels in practice: Trades and quotes. *The Econometrics Journal 12*(3), C1–C32. 41, 58, 64

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2009b). Realized kernels in practice: Trades and quotes. *Econometrics Journal 12*(3), C1–C32. 94, 110, 114

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2011). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics 162*(2), 149–169. 41, 58, 63, 64

Barndorff-Nielsen, O. E. and N. Shephard (2004b). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica 72*(3), 885–925. 40, 64, 74

Barsky, R. B. and L. Kilian (2001). Do we really know that oil caused the great stagflation? a monetary alternative. *NBER Macroeconomics annual 16*, 137–183. 17

Bartlett, M. (1963). The spectral analysis of point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 264–296. 100, 166

Baumeister, C. and L. Kilian (2016). Forty years of oil price fluctuations: Why the price of oil may still surprise us. *The Journal of Economic Perspectives 30*(1), 139–160. 15, 17, 20, 21

Beidas-Strom, S. and A. Pescatori (2014). *Oil Price Volatility and the Role of Speculation*. Number 14-218. International Monetary Fund. 23

Bessembinder, H., M. Panayides, and K. Venkataraman (2009). Hidden liquidity: an analysis of order exposure strategies in electronic stock markets. *Journal of Financial Economics 94*(3), 361–383. 36, 39, 40

Bhar, R. and D. Lee (2011). Time-varying market price of risk in the crude oil futures market. *Journal of Futures Markets 31*(8), 779–807. 38

Bohi, D. R. and M. A. Toman (1987). Futures trading and oil market conditions. *Journal of Futures Markets 7*(2), 203–221. 37

Bollen, N. P. and R. E. Whaley (2015). Futures market volatility: What has changed? *Journal of Futures Markets 35*(5), 426–454. 94

Budish, E., P. Cramton, and J. Shim (2013). The high-frequency trading arms race: Frequent batch auctions as a market design response. *Manuscript 6*. 39, 41, 42, 43

Buyuksahin, B., M. S. Haigh, J. H. Harris, J. A. Overdahl, and M. A. Robe (2008). Fundamentals, trader activity and derivative pricing. 26

Büyükşahin, B. and J. H. Harris (2011). Do speculators drive crude oil futures prices? *The Energy Journal*, 167–202. 14, 21

Cetin, U., R. Jarrow, P. Protter, and M. Warachka (2006). Pricing options in an extended black scholes economy with illiquidity: Theory and empirical evidence. *Review of Financial Studies 19*(2), 493–529. 89

Chamberlain, G. (1983). Funds, factors, and diversification in arbitrage pricing models. *Econometrica: Journal of the Econometric Society*, 1305–1323. 158, 162

De Charms, R. (2013). *Personal causation: The internal affective determinants of behavior*. Routledge. 43

Dovonon, P., S. Goncalves, and N. Meddahi (2013). Bootstrapping realized multivariate volatility measures. *Journal of Econometrics 172*(1), 49–65. 92, 110

Easley, D., T. Hendershott, and T. Ramadorai (2014). Leveling the trading field. *Journal of Financial Markets 17*, 65–93. 39

Epps, T. W. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association 74*(366a), 291–298. 58

Fattouh, B., L. Kilian, and L. Mahadeva (2012). The role of speculation in oil markets: What have we learned so far? 24, 26

Fattouh, B. and L. Mahadeva (2014). Causes and implications of shifts in financial participation in commodity markets. *Journal of Futures Markets34 34*(8), 757–787. 38

Fleming, J. and B. Ostdiek (1999). The impact of energy derivatives on the crude oil market. *Energy Economics 21*(2), 135–167. 22

Grossman, S. J. (1987). An analysis of the implications for stock and futures price volatility of program trading and dynamic hedging strategies. 52

Gsell, M. (2009). Algorithmic activity on xetra. *The Journal of Trading 4*(3), 74–86. 39

Gsell, M. and P. Gomber (2009). Algorithmic trading engines versus human traders-do they behave different in securities markets? 39

Guttentag, M. D. (2012). *8 Stumbling into crime: stochastic process models of accounting fraud*. Edward Elgar Publishing. 43

Hamilton, J. D. (2003). What is an oil shock? *Journal of econometrics 113*(2), 363–398. 17, 19

Hamilton, J. D. (2008). Understanding crude oil prices. Technical report, National Bureau of Economic Research. 34, 38

Hamilton, J. D. and J. C. Wu (2014). Risk premia in crude oil futures prices. *Journal of International Money and Finance 42*, 9–37. 26

Hart, H. L. A. and T. Honoré (1985). *Causation in the Law*. Oxford University Press. 43

Hayashi, T., N. Yoshida, et al. (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli 11*(2), 359–379. 40

Hedi Arouri, M. E. and D. Khuong Nguyen (2010). Oil prices, stock markets and portfolio investment: evidence from sector analysis in europe over the last decade. *Energy Policy 38*(8), 4528–4539. 38

Hellwig, M. (1996). Rational expectations equilibria in sequence economies with symmetric information: The two-period case. *Journal of Mathematical Economics 26*(1), 9–49. 52

Hendershott, T., C. M. Jones, and A. J. Menkveld (2011). Does algorithmic trading improve liquidity? *The Journal of Finance 66*(1), 1–33. 32, 39, 40

Hendershott, T. and R. Riordan (2009). Algorithmic trading and information. Technical report. 39

Hong, H. (2000). A model of returns and trading in futures markets. *The Journal of Finance 55*(2), 959–988. 95

Huang, R. D., R. W. Masulis, and H. R. Stoll (1996). Energy shocks and financial markets. *Journal of Futures Markets 16*(1), 1–27. 37

Jacod, J., Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter (2009a). Microstructure noise in the continuous case: the pre-averaging approach. *Stochastic Processes and their Applications 119*(7), 2249–2276. 41, 110

Jacod, J., Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter (2009b). Microstructure noise in the continuous case: The pre-averaging approach. *Stochastic Processes and their Applications 119*(7), 2249 – 2276. 94, 116

Kaufmann, R. K. (2011). The role of market fundamentals and speculation in recent price changes for crude oil. *Energy Policy 39*(1), 105–115. 23, 24

Kilian, L. (2008). The economic effects of energy price shocks. *Journal of Economic Literature 46*(4), 871–909. 22

Kilian, L. (2009). Oil price shocks, monetary policy and stagflation. 18, 22, 23, 24

Kilian, L. and B. Hicks (2013). Did unexpectedly strong economic growth cause the oil price shock of 2003–2008? *Journal of Forecasting 32*(5), 385–394. 23

Kilian, L. and D. P. Murphy (2014). The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics 29*(3), 454–478. 20, 23

Killian, L. (1998). Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics 80*(2), 218–230. 74, 75, 76

Killian, L. and U. Demiroglu (2000). Residual-based tests for normality in autoregressions: Asymptotic theory and simulation evidence. *Journal of Business and Economic Statistics 18*(1), 40–50. 74, 75

Kirilenko, A. A., A. S. Kyle, M. Samadi, and T. Tuzun (2014). The flash crash: The impact of high frequency trading on an electronic market. *Available at SSRN 1686004*. 51

Kirilenko, A. A. and A. W. Lo (2013). Moore's law versus murphy's law: Algorithmic trading and its discontents. *The Journal of Economic Perspectives 27*(2), 51–72. 42

Kyle, A. S. (1985a). Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society 53*(6), 1315–1335. 37

Kyle, A. S. (1985b). Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, 1315–1335. 38, 95

Lee, C.-C. and J.-H. Zeng (2011). Revisiting the relationship between spot and futures oil prices: Evidence from quantile cointegrating regression. *Energy Economics 33*(5), 924–935. 38

MacKinnon, J. G. (1992). Model specification tests and artificial regressions. *Journal of Economic Literature 30*(1), 102–146. 166

MacKinnon, J. G. (2002). Bootstrap inference in econometrics. *Canadian Journal of Economics/Revue canadienne d'économique 35*(4), 615–645. 145

MacKinnon, J. G. (2006). Bootstrap methods in econometrics. *Economic Record 82*(s1). 145

Moosa, I. A. and N. E. Al-Loughani (1995). The effectiveness of arbitrage and speculation in the crude oil futures market. *Journal of Futures Markets 15*(2), 167–186. 37

Muirhead, R. J. R. J. (1982). Aspects of multivariate statistical theory. Technical report. 100, 105, 107, 109, 118, 119, 122, 123, 129, 130, 132, 135, 137, 141, 142, 146, 149, 152, 161, 172

Nakajima, K. and K. Ohashi (2012). A cointegrated commodity pricing model. *Journal of Futures Markets 32*(11), 995–1033. 38

Newey, W. K. and K. D. West (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. 12

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics 92*(4), 1004–1016. 161, 162

online (2016). https://www.macrotrends.net/1369/crude-oil-price-history-chartr kernel description. 16, 19

Overdahl, J. A. (1987). The use of crude oil futures by the governments of oil-producing states. *Journal of Futures Markets 7*(6), 603–617. 37

Peroni, E. and R. McNown (1998). Noninformative and informative tests of efficiency in three energy futures markets. *Journal of Futures Markets 18*(8), 939–964. 37

Quan, J. (1992). Two-step testing procedure for price discovery role of futures prices. *Journal of Futures Markets 12*(2), 139–149. 37

Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial management review 6*(2), 41–49. 95

Schaffer, J. (2010). Contrastive causation in the law. 43

Silvério, R. and A. Szklo (2012). The effect of the financial sector on the evolution of oil prices: Analysis of the contribution of the futures market to the price discovery process in the wti spot market. *Energy Economics 34*(6), 1799–1808. 38

Stiglitz, J. E. (2002). Information and the change in the paradigm in economics. *American Economic Review 92*(3), 460–501. 52

Switzer, L. N. and M. El-Khoury (2007). Extreme volatility, speculative efficiency, and the hedging effectiveness of the oil futures markets. *Journal of Futures Markets 27*(1), 61–84. 38

Voev, V. and A. Lunde (2007). Integrated covariance estimation using high-frequency data in the presence of noise. *Journal of Financial Econometrics 5*(1), 68–104. 74

Wang, T., J. Wu, and J. Yang (2008). Realized volatility and correlation in energy futures markets. *Journal of Futures Markets 28*(10), 993–1011. 38

Williams, J., P. Dovonon, and A. Taamouti (2017). testing the number of factors in high frequency data. Working paper, Durham University. 168, 185

Wright, R. W. (1985). Causation in tort law. *California Law Review 73*(6), 1735–1828. 43

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics 160*(1), 33–47. 41